Title

Learning Through Comparisons: An Evaluation of Simultaneous Comparison Trials in Perceptual Category Learning

Author

Jacoby, Victoria L

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Learning Through Comparisons:

An Evaluation of Simultaneous Comparison Trials in Perceptual Category Learning

A dissertation submitted in partial satisfaction

of the requirements for the degree Doctor of Philosophy

in Psychology

by

Victoria Leigh Jacoby

2024

ABSTRACT OF THE DISSERTATION

Learning Through Comparisons:

An Evaluation of Simultaneous Comparison Trials in Perceptual Category Learning

by

Victoria Leigh Jacoby

Doctor of Philosophy in Psychology

University of California, Los Angeles, 2024

Professor Philip Kellman, Chair

Whether it be your ability to consistently recognize the face of your friend across varying contexts, or the ability of a dermatologist to differentiate cancerous skin lesions from benign ones, many tasks and domains are reliant upon the ability to classify items into one of many possible categories. This can be challenging, especially when category members present in diverse ways or closely resemble members of other categories. Research has shown that category acquisition can be facilitated by opportunities to directly compare items together. In particular, comparisons may support rapid improvements in the discovery and pick up of information, resulting in deeper processing of category structure and enhanced category representations. The goal of this dissertation was to elucidate the most effective methods for constructing

comparisons to accelerate the mechanisms of perceptual learning that underlie successful categorization. In particular, emphasis was given to exploring the value of *paired comparison trials*, which involved the concurrent presentation of items from two different categories for discrimination.

In Experiments 1 and 2, learning trial structure and task were manipulated to test paired comparison learning against more common classification-based approaches. Results revealed that paired comparison learning was an effective way to learn the classification of a large set of categories, particularly when the domain was novel to the learner. Experiment 3 broke down paired comparisons into its separate learning components to evaluate how learning is advanced. Results revealed asymmetric learning gains in favor categories framed as the target of the trial relative to categories framed as the distractor. Experiment 4 measured the perceptual changes induced by paired comparison learning, finding evidence for within-category compression and between-category expansion. Finally, Experiment 5 evaluated adaptive learning methods to enhance comparison efficacy by testing an adaptive comparison procedure, previously used in face learning, in the domain of skin lesion classification. A partial replication of results was observed, and suggestions are given as to how the procedure may be improved. Altogether, this work has important implications for our understanding of the role of perceptual learning in high-level tasks generally, as well as for how it interacts with comparison opportunities specifically.

The dissertation of Victoria Leigh Jacoby is approved.

Elizabeth Ligon Bjork

Sally J. Krasne

Hongjing Lu

Christine M. Massey

Philip Kellman, Committee Chair

University of California, Los Angeles

2024

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 2

Chapter 3

Chapter 4

Chapter 5

LIST OF TABLES

Chapter 4

ACKNOWLEDGMENTS

I would like to acknowledge and thank the members of the UCLA Human Perception Lab for all of their help in the research reported in this dissertation, as well as for shaping my graduate school experience. To Tim, thank you for sharing your programming expertise, your quick problem solving, and your detailed explanations. To Austin, thank you for the laughs shared in our office and for bearing with me on my more sleep-deprived days. To Christine Massey, thank you for all of your careful and thoughtful feedback over the years and for making me feel comfortable and welcome here from the very beginning. And to my advisor Phil Kellman, thank you for guiding and inspiring me and my research. I am deeply appreciative of the patience, respect, and encouragement you have given me throughout my time in this program.

I would also like to acknowledge and thank my friends and family for all they have done and continue to do for me. To my friends, thank you for cheering me on, checking in on me, and letting me vent to you (more than a normal person ever should) about the number of emails I need to reply to on any given day. To my parents, thank you for always pushing me to strive for my best and for always believing in my ability to do whatever I set my mind to. You have done so much to put me in a position where I could dedicate my time and energy to pursuing an education at this level, and I will always be grateful for that.

Finally, to my beautiful wife Ariana, thank you for supporting me in practically every way a human being can possibly be supported.

Portions of this research were supported by National Institutes of Health grant R01CA236791.

Chapter 4, Experiment 3 is a modified version of:

Jacoby, V.L., Massey, C.M., & Kellman, P.J. (in press). Target vs. Distractor: Does the role of a category in comparisons influence learning? Evidence from skin cancer classification. *Proceedings of the Annual Meeting of the Cognitive Science Society*.

A portion of the research reported in Chapter 5 is a modified version of:

Jacoby, V.L., Massey, C.M., & Kellman, P.J. (under review). Adaptively triggered comparisons enhance perceptual category learning: Evidence from face learning.

VITA

EDUCATION

University of California, Los Angeles
M.A. in Psychology (awarded December 2019)

Arizona State University
B.S. in Psychology; Concentration in Psychological Science (awarded May 2018)
Honors: *Summa Cum Laude*

PUBLICATIONS

Jacoby, V.L., Massey, C.M., & Kellman, P.J. (in press). Target vs. Distractor: Does the role of a category in comparisons influence learning? Evidence from skin cancer classification. *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Jacoby, V. L., Massey, C. M., Mettler, E., & Kellman, P. J. (2022). Comparisons in Adaptive Perceptual Category Learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 44, No. 44).

Kellman, P. J., Jacoby, V., Massey, C. M. & Krasne, S. (2022). Perceptual learning, adaptive learning, and gamification: Educational technologies for pattern recognition, problem solving and knowledge retention in medical learning. In: M. Lee & H. Witchel (Eds.), *Technologies in Biomedical and Life Science Education: Approaches and Efficacy for Learning.* American Physiological Society *Methods in Physiology Series.* New York: Springer.

MANUSCRIPTS IN PROGRESS OR UNDER REVIEW

Jacoby, V.L., Massey, C.M., & Kellman, P.J. (under review). Adaptively triggered comparisons enhance perceptual category learning: Evidence from face learning.

Jacoby, V.L., Massey, C.M., & Kellman, P.J. (in progress). Learning through visual contrasts: Paired comparisons in facial identification and skin cancer classification.

CONFERENCE PRESENTATIONS

Jacoby, V. L.*, Massey, C.M., & Kellman, P.J. (2024, May). *Comparison Training Improves Perceptual Learning of Skin Cancer Diagnoses.* Poster presented at the 24th Annual Meeting of the Vision Sciences Society, St. Pete Beach, FL, United States.

Jacoby, V. L.*, Massey, C.M., & Kellman, P.J. (2023, November). *Perceptual Learning Based Entirely on Comparisons is Best: Evidence from Skin Cancer Classification.* Poster presented at the 64th Annual Meeting of the Psychonomic Society, San Francisco, CA, United States.

Jacoby, V. L.*, Massey, C.M., & Kellman, P.J. (2023, May). *Paired Comparisons Effectively Drive the Learning of Multi-Category Perceptual Learning*. Poster presented at the 23rd Annual Meeting of the Vision Sciences Society, St. Pete Beach, FL, United States.

Jacoby, V.L.*, Massey, C.M., Mettler, E., & Kellman, P.J. (2022, November). *Paired Comparisons in Perceptual Category Learning*. Poster presented at the 63rd Annual Meeting of the Psychonomic Society, Boston, MA, United States.

Jacoby, V.L.*, Massey, C.M., Mettler, E., & Kellman, P.J. (2022, July). *Comparisons in Adaptive Perceptual Category Learning*. Poster presented at the 44th Annual Meeting of the Cognitive Science Society, Toronto, ON, Canada.

Kellman, P. J., Jacoby, V*., Massey, C. (2021, November). *Enhancing Perceptual Learning Through Adaptive Comparisons*. Paper presented at the Virtual 62nd Annual Meeting of the Psychonomic Society.

Jacoby, V. L.*, Massey, C. M., & Kellman, P. J. (2021, May) *Enhancing Perceptual Learning Through Adaptive Comparisons*. Poster presented at the Virtual 21st Annual Meeting of the Vision Sciences Society.

Brewer, G. A.*, Jacoby, V., Pitaes, M., & Vogel, A. (2019, June) Biased Judgements Regarding Other People's Prospective Memory Failures. In L. Kvavilashvili (Chair), *Remembering to execute future intentions: Latest findings in prospective memory research* [Symposium]. The Society for Applied Research in Memory and Cognition Conference, Cape Cod, MA, United States.

*Indicates Presenting Author

TEACHING EXPERIENCE

2018 - 2024
Teaching Assistant/Associate/Fellow, *UCLA*

Courses: Cognitive Psychology, General Psychology Laboratory, Introduction to Cognitive Science, Introductory Psychology, Laboratory in Cognitive Psychology, Principles of Behavioral Neuroscience, Research Methods, Sensation and Perception

CHAPTER 1

General Introduction and Dissertation Goals

Categorization enables the meaningful grouping of entities that share common attributes, whether in definition, function, meaning, structure, or form. This organization is crucial for managing the vast amount of information we encounter daily and can be used to guide our interactions, enable more efficient decision making, and assist us in navigating our environment. Examples of this can be seen in everyday situations, such as needing to determine whether the new object sitting on your kitchen counter is a muffin or a sponge and subsequently using that information to determine whether or not you should eat it. Categorization is also integral in highly specialized tasks, such as a dermatologist investigating whether a skin lesion shows visual signs of malignancy and making the decision as to whether it should be biopsied.

Critically, the formation of categories allows us to leverage our accumulated knowledge and experiences to assist in making decisions about newly encountered objects or information. This includes using previously learned conceptual and semantic information about category members, such as the dermatologist's knowledge that a certain type of skin cancer is more frequently found on sun-exposed areas of the body than others, as well as recognizing implicitly learned perceptual patterns, such as the way different skin lesions may subtly vary in symmetry, color, and texture.

Given the ubiquitous nature of categories, the question arises: how can we most effectively prepare learners to discern whether a novel entity belongs to one category or another? Extensive research has suggested that direct comparison of items can play an important role in facilitating category acquisition (e.g., Kurtz & Gentner, 2013; Spalding & Ross, 1984). Categories seldom exist in isolation (e.g., the category of cancerous skin lesions is often

considered in relation to the category of non-cancerous skin lesions), and by intentionally aligning critical features and patterns across different stimuli, learners can more readily identify the common characteristics that unite members of the same category and the information that distinguishes them from others. In the present dissertation, I examine the significant relationship between the perceptual processes underlying categorization and the role of comparisons in enhancing them.

Perceptual Category Learning

Perceptual categories have been broadly described as a collection of similar objects united by common perceptual features, patterns, or rules (Ashby & Maddox, 2005; Mandler, 1997). This type of category is often differentiated from conceptual categories or concepts, which involve groupings based on more abstract, non-observable ideas. For example, *types of birds* may be considered a perceptual category, as one may learn to distinguish and classify different types of birds based on perceptual information alone (e.g., size, vocalizations, feather pattern). This can be contrasted with concepts such as *justice* or *living things*, which may be informed by observable perceptual information, but ultimately involve more abstract processing of the information (e.g., what does it truly mean for something that moves and acts like it's alive to actually be alive? What is the essence of *life*?). Importantly, this does not imply that perceptual categories cannot contain any conceptual information or that concepts must be entirely amodal, as it is often the case that both types of information interact (Sloutsky, 2010). Rather, regardless of whether conceptual information may be present, the acquisition of perceptual categories is dependent primarily upon the perceptual information available. In the

present work, we focus specifically on perceptual categories whose members share a common perceptible visual organization.

The acquisition of perceptual categories relies on learning the distinguishing features among items of different categories and identifying the features that unite members within the same category (Gibson, 1969; Homa & Chambliss, 1975). These categories can be further classified as being either well-defined or ill-defined. A well-defined category, also termed a rule-based category, possesses clear dimensions and specific membership rules, accommodating limited variability. Such categories vary on critical features along one or more dimensions, with categorization rules often explicitly identifiable and verbalizable after training. In contrast, ill-defined categories lack clear boundaries between categories and necessitate information integration for classification. Acquisition of these categories requires the learner to incorporate changes from multiple dimensions simultaneously at a pre-decisional stage to discern a type of "optimal rule" for making subsequent classification judgments (Ashby et al., 1998). In an experimental setting, stimuli like abstract forms and patterns are also often created and modified to constitute ill-defined categories (Medin & Schaffer, 1978), though it has been argued that most naturally occurring categories are also ill-defined (Neisser, 1967; Rosch, 1973; Wittgenstein, 1953).

Some models of categorization have viewed the learning of perceptual categories as entirely memory-dependent. In this view, the perceptual input received is often treated as fixed, such that categories are perceived to have the same constant features and dimensions from initial exposure to mastery. Categorization judgements are then made by comparing the similarity of the currently viewed stimuli to a stored category prototype (Homa, 1984; Posner & Keele, 1968, 1970; Reed, 1972), to all previously seen exemplars (Medin & Schaffer, 1978; Nosofsky, 1986,

1988), or to some combination of both (Smith & Minda, 1998). However, there is abundant evidence that perception is not fixed, but rather changes dynamically to optimize task performance (e.g., Folstein et al., 2013; Goldstone et al., 2001; Hock et al., 1987). In particular, new features and patterns contained within a category can be discovered and emphasized, and irrelevant perceptual information can be ignored. These are not conscious processes; rather, selective tuning of perceptual mechanisms to deliver relevant features and relations can become automatic (e.g., Schneider & Shiffrin, 1977). In recent years, neural networks and deep learning models have given clear form to the way selective mechanisms might evolve through learning that gradually strengthens connection weights between inputs and outcomes (Rumelhart, 1989; Goodfellow et al., 2016). Category acquisition and subsequent categorization is then no longer solely dependent upon the memorization and subsequent matching of these features, but rather in the implicit ability to find and use the most critical information. These changes in perception are the result of perceptual learning— broadly defined as experience-driven improvements in the pickup of information (Gibson, 1969).

Though much of the contemporary perceptual learning literature has focused on investigating these improvements in simple stimuli that vary across a single dimension, such as in the discrimination of spatial frequency (Bennett & Westheimer, 1991) or stimulus orientation (Dosher & Lu, 1999; Vogels & Orban, 1985), perceptual learning has also been shown to be vital in tasks involving more complex stimuli, such as in the classification of butterflies (Mettler & Kellman, 2014), the discrimination of faces (Mundy et al., 2007), or the interpretation of medical images (e.g., Kellman et al., 2023; Krasne et al., 2013; Marris et al., 2023; Roads et al., 2018). Eleanor Gibson, who pioneered the field of perceptual learning (Gibson, 1969) focused on examples such as interpreting thermal imagery, distinguishing aircraft types, or normal reading.

However, in the resurgence of perceptual learning research beginning around the early 1990s, the focus was often on basic acuities, such as orientation or motion direction discrimination, largely based on the hypothesis that these changes might reveal plasticity in early cortical receptive fields. That hypothesis has largely given way to views of perceptual learning that emphasize selective mechanisms that may operate across a variety of tasks and levels (Ahissar & Hochstein, 2004; Dosher et al., 2013; Garrigan & Kellman, 2008; Gilbert et al., 2009; Petrov et al., 2005; for discussion, see Kellman & Garrigan, 2009).

It has been argued that there are two main types of changes in perception that can occur as a result of perceptual learning: *discovery* and *fluency* effects (Kellman, 2002). Discovery effects refer to the perceiver's ability to find and amplify information within the presented stimuli that is relevant to the task at hand. In the context of perceptual category acquisition, perceptual learning enables the discovery and preferential extraction of diagnostic features common to members of the same category as well as the distinguishing features of different categories. Fluency effects then refer not to the ability to find information, but to the efficiency with which the perceiver can extract this information. This could include extracting information more quickly and automatically (Kellman & Garrigan, 2009: Shiffrin & Schneider, 1977), as well as extracting and storing larger "chunks" of information at a time (Chase & Simon, 1973).

Perceptual learning processes may be gradual and implicit, often requiring extensive practice or experience with relevant instances to see significant changes in perception. Given this, there is significant value to finding ways to structure practice and learning events to optimize the opportunities for perceptual learning to occur. In the present dissertation, I investigate how the incorporation of comparisons and adaptive learning technology in training may enhance perceptual learning and the acquisition of complex perceptual categories.

Comparisons

A considerable amount of research has been dedicated to exploring the role of comparison in learning and cognition generally (e.g. Gentner & Markman, 1997; Medin et al., 1993), as well as in improving category acquisition and transfer in particular. Most simply, comparisons involve providing the opportunity to look for similarities and/or differences between items, often achieved by presenting stimuli concurrently. In the context of learning perceptual classifications, the mechanisms underlying perceptual learning in categorization can directly benefit from comparison opportunities.

Gibson (1969) proposed that much of perceptual learning involved the brain's search for *distinguishing features*. Rather than compile ever better descriptions of members of categories, the focus in learning seems to be discovery of the stimulus attributes that make the difference for membership in contrasting categories, a proposition rooted in work on speech perception (Jacobson & Halle, 1956) and supported by some evidence in visual perception (Gibson et al., 1962). This suggests that mechanisms of perceptual learning are deeply rooted in supporting category learning. Finding what differentiates a category from another cannot be achieved just by looking at one category alone; distinguishing features are relational, requiring both categories to be considered. While some have defined comparison to be its own learning mechanism, this dissertation will define comparison as, on the experimenter or teacher's part, an independent variable involving the methods of presentation used in learning, and on the learner's part, a process or strategy integrated into learning that interacts with the mechanisms of perceptual learning. In other words, comparisons alone do not advance learning, but rather comparisons provide the opportunity for more rapid and complete differentiation or discovery of perceptual patterns which can then advance learning.

Experimental results have provided support for the role of comparison to improve differentiation for presented items (Mundy et al., 2007, 2009), and consequently modify and shape the perception and representation of each item or category (Medin et al., 1993; Spalding & Ross, 1984). Further, relative to processing each item independently, the comparison of categories can lead to deeper processing, such that participants shift focus from individual surface features to the commonalities and differences in overall structure (Goldstone et al., 2010).

Despite the compelling evidence for comparison as a tool to advance learning, how to best integrate comparisons in learning is not clearly established. There are several elements that may influence the effectiveness of comparison. These can be elements of the experimental design, as well as elements of the stimuli/categories themselves. While the primary requirement is the ability to compare two or more items, the type of items that should be compared, and how one should interact with the presented items, has varied greatly across studies.

An overwhelming focus of category comparison research, both in natural and artificial domains, has been on manipulating the content included in a comparison, namely whether items being compared are from the same or different categories, (e.g., Andrews et al., 2011; Higgins & Ross, 2011; Kang & Pashler, 2012; Kok et al., 2013). This has led to further research focusing on how the items best suited for comparison may interact with characteristics of the presented items and category boundaries, including the similarity of the presented items, the similarity of one category to other categories in a learning set, and the amount of variability present within the same category (Carvalho & Goldstone, 2014; Jee et al., 2013). In recent work, we have extended this to consider how the items best suited for comparison may also differ across each learner (Jacoby, Massey, & Kellman, 2021).

While the question of which types of items should be presented in a comparison is important, insufficient consideration has been given to how these items should be displayed and what participants should be instructed to do with them throughout training. For example, there is some debate as to just how many items should be presented at a time for a comparison to be most effective. While some approaches to comparison have shown a benefit of presenting two or more items side-by-side in learning (Andrews et al., 2011; Homa et al., 2014, Jee et al., 2013) others have suggested that effective comparisons can be promoted in the successive presentation of only one single stimulus at a time (e.g., Carvalho & Goldstone, 2015ab, 2017; Kang & Pashler, 2012).

Further, the effectiveness of a comparison may differ based on how a participant is required to interact with the presented stimuli throughout learning. Previous attempts to study comparison have utilized a variety of approaches including instructing participants to passively view stimuli (e.g., Kang & Pashler, 2012; Kornell & Bjork, 2008), make same-different judgements between presented stimuli (e.g., Angulo et al., 2019; Higgins & Ross, 2011), as well as classify presented items (e.g., Andrews et al., 2011; Carvalho & Goldstone, 2014; Homa et al., 2014). While prior research has begun to contrast passive and active approaches (e.g., Carvalho & Goldstone, 2015a; Levering & Kurtz, 2015; Patterson & Kurtz, 2020), little research has focused on how different active approaches may interact differently with learning.

Adaptive Learning

Recently, adaptive learning methods have also been incorporated into perceptual category learning paradigms in an attempt to advance learning. Adaptive learning refers to methods and technologies that seek to personalize and optimize the learning process for an individual by

adjusting events in learning based on the individual's responses or performance. Approaches can range from relatively simple, such as using pre-testing to establish a learner's starting point, to much more complex, such as the continuous tracking of performance to continually decide the sequence of learning events. Moreover, adaptive methods are most effective when they are based on learning principles that have already been shown to meaningfully influence learning.

One such example is an adaptive approach to spacing that uses participant performance to schedule item presentations called the Adaptive Response Time-Based Sequencing System (ARTS) (Mettler et al., 2011, 2016). Substantial research in the learning sciences has reported a benefit of spacing, as opposed to massing, of the to-be learned material, often referred to as the *spacing effect* (Bjork & Bjork, 2011; Karpicke & Roediger, 2007). Ideally, practice with a given problem or item should be repeated at the point in which recalling information is at its most difficult, but has not yet been forgotten (Bjork & Bjork, 1992, 2011; Mettler & Kellman, 2014). Most work on spacing, however, has used predetermined schedules. Yet the difficulty of an item for perceptual classification likely varies across individuals, items, and different times within a learning session, making predetermined schedules of spacing inherently suboptimal. Without monitoring an individual's learning strength for a given item, it is impossible to determine the optimal amount of spacing (Mettler et al., 2016).

Adaptive learning systems have existed for some time (e.g., Atkinson, 1974). Although there have been many variations, until recently almost all systems have relied exclusively on accuracy of learner performance to determine entry levels, spacing and sequencing in learning, and mastery or advancement. Accuracy is one important indicator of learning strength, but by itself is limited. An accurate response can be quick, effortless, and confident, or slow, effortful, and uncertain (and many combinations in between). Experimental evidence has demonstrated

that response times can be used to infer learning strength (Benjamin & Bjork, 1996; Pyc & Rawson, 2009), such that lower response times are indicative of greater learning strengths. Notably, this applies only to response times for accurate responses, as quick and effortless incorrect responses do not inherently suggest any benefit over slow or uncertain incorrect responses. Adaptive learning systems can better assess underlying learning strength by adding response time measurements to accuracy.

The ARTS system was developed to incorporate response times to better gauge learning strength, and use those measurements to guide both spacing of learning items and mastery (Kellman, 2002; Mettler, Kellman & Massey, 2011, 2016). ARTS has been shown to enhance factual learning better than a number of alternative approaches, including other adaptive learning systems (Mettler, Massey & Kellman, 2016). When applied to perceptual learning, the ARTS system has also been shown to increase learning efficiency (Mettler & Kellman, 2014).

Some prior research has suggested that adaptive methods can effectively be applied to the use of comparison trials in learning (Thai et al., 2015), and recently, a new adaptive approach to comparisons, *Adaptively Triggered Comparisons*, has been introduced and used in conjunction with the ARTS system (Jacoby, Massey, & Kellman, 2021). Instead of only monitoring whether a category was classified correctly or incorrectly, this system pays attention to the specific category label chosen on incorrect trials. By monitoring patterns in participant classification errors, a learner's confusions between categories can be made evident. For example, if Category A is repeatedly labeled as Category B and/or Category B is repeatedly labeled as Category A, then not only do we know learning strength is low for both categories A and B, but also that there is a confusion between them. When learning errors indicate such a confusion, an adaptively-triggered comparison (ATC) trial is presented, in which one exemplar from each of

the two confused categories is presented simultaneously for discrimination. This triggered intervention then allows for a direct comparison of categories that can lead to the extraction of critical invariants. This, in theory, will promote encoding of distinguishing information for each item, such that the next time a participant encounters either of those categories, they will be less likely to confuse them. When applied to a face learning paradigm, the presence of ATC trials improved learning efficiency relative to conditions in which no simultaneous comparisons were used, as well as to a condition with an equal number of non-adaptive simultaneous comparisons, suggesting that using learner performance to determine the contents and timing of comparisons can enhance perceptual category learning.

Goals of This Dissertation

In this dissertation, I investigate how comparisons may be used to enhance the learning of complex perceptual categories. First, I break down the elements involved with different approaches to comparison to advance our understanding of how they may influence comparison efficacy and perceptual learning. Chapter 3 investigates whether the task of a comparison trial plays a meaningful role in learning outcomes. Here, paired comparison trials that emphasized the visual discrimination of a pair of items are considered in contrast to more common classification-focused trials in which learners were presented with one or two instances and required to classify each item into a large set of potential categories. Experiments 1 and 2 tested paired comparison trials across two separate, real-world domains. Additional consideration is given to the value of utilizing a simultaneous presentation of items relative to a sequential presentation. Chapter 4 contains two experiments that aimed to more closely investigate how learning progresses through paired comparison training. In Experiment 3, the paired comparison trial format was broken

down into separate components to evaluate what is learned about each presented category on a given trial, as well as how that may influence the overall learning of all categories in a set. Experiment 4 characterized how perception changes as a result of comparison-based learning by measuring changes in category representations. Finally, in Chapter 5, the value of adaptive learning technology in perceptual categorization tasks is tested by expanding upon adaptive approaches to comparison. I start by reviewing our preliminary work on adaptive comparisons in face identification, before introducing Experiment 5, which aimed to replicate and expand upon the original ATC research in a new domain: skin lesion classification.

CHAPTER 2

Learning Domains

The majority of research on comparisons in category learning tasks has been conducted with artificial stimuli – created by the experimenter for the purpose of studying categorization. As a result, these stimuli are often simple, abstract, and highly controlled with clear rules for membership and definitive category boundaries. Artificial stimuli categories have included things like dot patterns or line drawings, as well as more detailed, computer-generated forms with consistent overall structure but varied features. However, given the importance of multi-category perceptual classification to many real-world educational and training applications, there has been an increased interest in evaluating learning with comparisons in complex, naturalistic domains. This has included domains such as fingerprint matching (Searston & Tangen, 2017), geologic classification of rocks (Meagher et al., 2017), and medical image interpretation (Evered et al., 2013; Kok et al., 2013; Marris et al., 2023; Sha et al., 2020; Sukut et al., 2023). These domains welcome greater variety that may be more representative of the characteristics of most categories encountered in the world.

A primary focus of this dissertation is to understand the potential benefit of comparison in meaningful real-world contexts; thus, the present work investigated comparisons in the context of two different, naturalistic domains: human face perception and the differential diagnosis of skin lesions. Although these domains differ in many ways, they are united in that the acquisition of categories (faces or skin lesions) does not depend on learning a single critical feature or rule, but rather is dependent upon some level of information integration or complex pattern extraction. Similarly, both domains are characterized by a large number of distinct categories, each with many examples available for training and testing generalization of learning.

Background and General Methods

Human Face Perception

Background

Face perception is a critical part of everyday life for most people. The ability to recognize the face of your friend, as well as differentiate it from the face of your coworker, relies on our ability to recognize a face under a wide range of conditions. This can include changes in the environment, such as variation in lighting and viewpoint, changes made on or around the face, such as the addition of cosmetics or accessories, and even changes in the face itself, such as muscle movements or different expressions. Due to all of this variation, outside of static representations (like photographs), the perceptual information received from looking at the same face on two different occasions will almost never be identical.

Given the importance of face recognition for social interactions, humans may have evolved innate perceptual foundations for face perception, with evidence showing a preference for faces even in newborn infants (Goren et al., 1975; Morton & Johnson, 1991). Additionally, our consistent, repeated exposure to faces further leads to the development of face processing expertise. However, that is not to say our face perception is perfect or unable to be changed. Studies have demonstrated that perceptual learning training can still be used to effectively enhance face perception in adults with normal face processing abilities (e.g., Hussain et al., 2009; Pachai et al., 2011), as well as those afflicted with face-perception deficits (Corrow et al., 2019; Davies-Thompson et al., 2017).

In the work presented here, learning to recognize a face refers to learning facial identity; that is, becoming able to recognize the same person from varying views. Some, but not all, research on perceptual learning in face perception has used this characterization, where a label

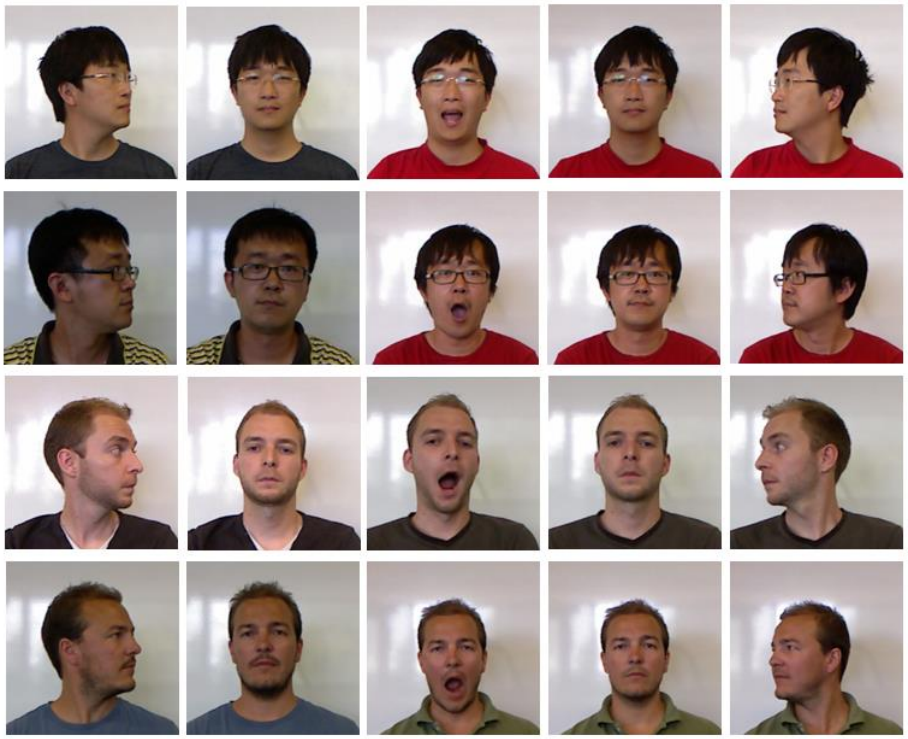
(typically a name) refers to a person or face category rather than a specific image. Other research (e.g., Gold et al., 2004) has used single images of each person. These choices often relate to particular experimental goals. Use of a single image for each face to be learned, however, is open to the concern that the learning involves memorization of each particular image and association with a particular response, rather than the extraction of structure from variable views, which may include different vantage points, clothing, and other non-face details. The latter type of task both more clearly involves perceptual learning and also is ecologically valid in terms of what face (person) recognition requires in natural circumstances.

Database

The stimuli used to study face categorization in the present work were taken from the EURECOM Kinect Face Database (Min et al., 2014). This dataset contains images of 52 individuals (14 females, 38 males) with each person photographed under 9 different conditions. Additionally, each of these conditions are repeated twice, with the second iteration being captured on average 1 to 2 weeks later. As a result, most individuals are photographed with differences in their non-face details (e.g., clothing, hairstyle). Images were 256 X 256 pixels in size.

In the experiments reported here, we used a subset of this database containing 22 individuals (all male) photographed across 4 different image conditions. These included a neutral expression, mouth opened pose, left profile, and right profile. Background and final image size are held constant across all exemplars. Unlike many other face datasets which remove all non-face details, the faces remain unaltered, thus yielding higher ecological validity. Example of face-categories from this dataset are shown in Figure 1. Our preliminary work in adaptively triggered comparisons and Experiment 1 were conducted using these stimuli.

Figure 1
Exemplars for Face Categories



Note. Each row in this figure depicts the exemplars for one face category. All face images used in this work are publicly available through the EURECOM Kinect Face Dataset (Min et al., 2014).

Dermatology: Skin Lesion Classification

Background

In addition to being involved in everyday life, perceptual categories are also a vital part of many highly specialized fields. In many medical imaging domains, the detection of abnormalities often requires the discovery and interpretation of complex visual patterns. Cancer detection and lesion classification in dermatology is one such domain, where perceptual classifications of trained experts guide important medical outcomes.

Cancers of the skin are by far the most common of all types of cancer, and of these by far the most common are basal and squamous cell skin cancers with about 5.4 million being diagnosed each year in about 3.3 million Americans (*American Cancer Society*, 2023a). Melanoma, the third most common skin cancer, is the leading cause of skin cancer mortality (Islami et al., 2021). The American Cancer Society estimates that, in 2023, about 97,610 new melanomas will be diagnosed in the US and about 7,990 people will die of melanoma (*American Cancer Society*, 2023b). From 2016-18, the average annual number of adults treated for any skin cancer was 6.1 (95% CI: 5.6, 6.6) million and the overall estimated annual cost of treatment alone was $8.9 billion (Kao et al., 2023). Average number of Years of Potential Life Lost per death are approximately 15 for melanoma and 10 for non-melanoma skin cancer (NMSC). Costs attributable to melanoma and NMSC for morbidity were estimated at $39.2 million and $28.9 million, and $3.3 billion and $1.0 billion for mortality (e.g., lost income due to premature death) (Guy & Ekwueme, 2011).

Given the severe costs and consequences associated with cancers of the skin, it is imperative that experts are equipped with the best training possible to increase detection rates of dangerous carcinomas and melanomas. While increasing declarative knowledge and explicitly learned membership rules can provide valuable assistance in certain instances of perceptual classification, expertise in visual discrimination is largely dependent upon more implicit processes of perceptual learning (Gibson, 1969; Kellman & Garrigan, 2009). In medical image interpretation in particular, the central role of perceptual learning has been increasingly recognized in recent years (Waite et al., 2019; Guegan et al., 2021). In the case of skin cancer detection, a learner may be informed about which features (e.g., size, shape, symmetry, color) are most common in cancerous versus benign lesions; however, because many of these features

vary across lesions of the same category, as well as appear very similarly in lesions of different categories, this declarative knowledge alone is insufficient.

Recent work has sought to introduce perceptual learning interventions across multiple visually-rich medical domains to help accelerate the development of perceptual expertise (for a review, see Kellman et al., 2022). In dermatology in particular, perceptual learning modules designed to improve the classification of skin lesion categories successfully accelerated category acquisition and transfer for medical students (Rimoin et al., 2015) and novice learners (Kellman et al., 2023).

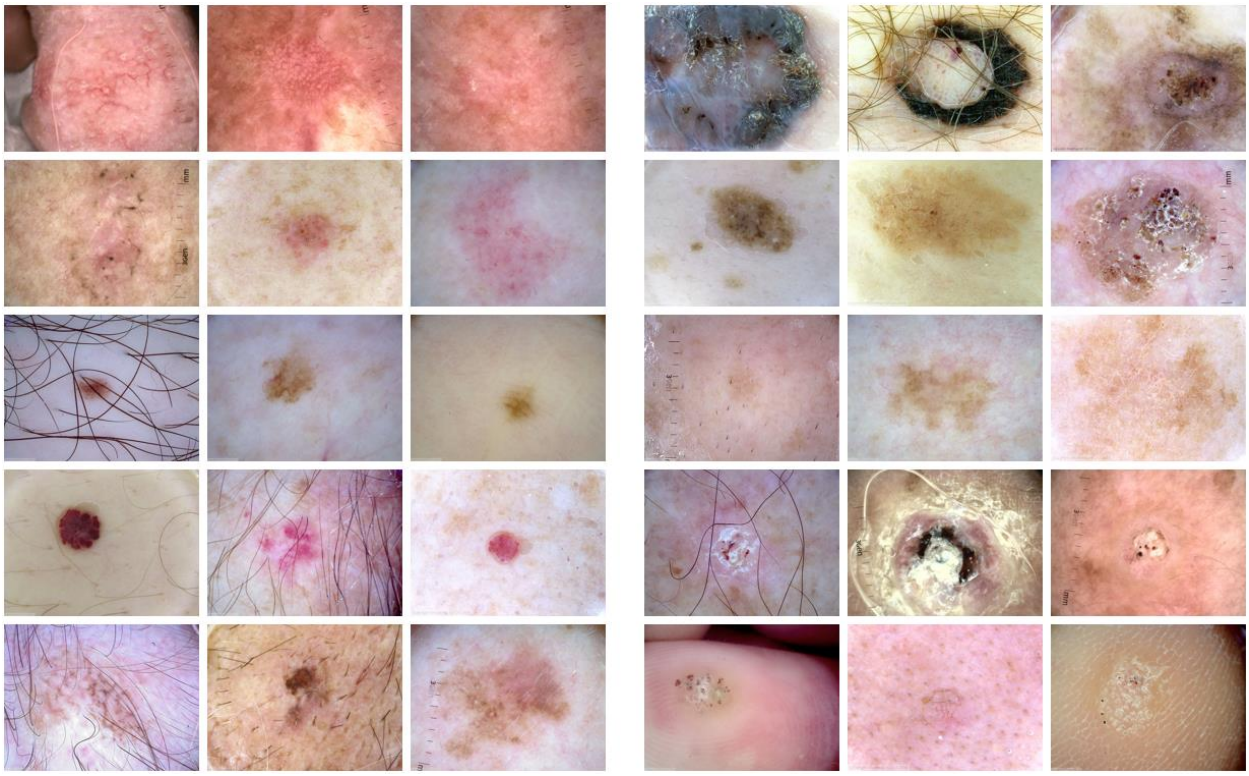
Database

The skin lesion images included in the present work were obtained from the MoleMap Database. This database contains approximately 2.5 million lesion images, with image verification completed by a combination of expert dermatologists, melanographers, and AI systems. A subset of 10 different categories of skin lesions were used in the present experiments including: Actinic Keratosis, Basal Cell Carcinoma, Benign Nevus, Haemangioma, Lentigo Maligna Melanoma, Nodular Melanoma, Seborrheic Keratosis, Solar Lentigo, Squamous Cell Carcinoma, and Wart. This particular subset of images was chosen on the basis of original dermatologic diagnosis, verification via biopsy when appropriate, and good image quality. Each category contained 20-100+ distinct presentations. All images are available in both clinical and dermoscopic[1] views; however, only the latter was used in the work reported here. Images were 1600 x 1200 pixels in size. An example of these stimuli can be seen in Figure 2. Experiments 2-5 were conducted using subsets of categories and exemplars from this database.

[1] Dermoscopic views are magnified and acquired using light shining at an angle to avoid reflections from the surface, with the result that deeper levels of the epidermis can be imaged.

Figure 2

Exemplars for Skin Lesion Categories



Note. Three images from each of the 10 skin lesion categories are presented above. The left column, from top to bottom, depicts: Actinic Keratosis, Basal Cell Carcinoma, Benign Nevus, Haemangioma, Lentigo Maligna Melanoma. The right column, from top to bottom, depicts: Nodular Melanoma, Seborrheic Keratosis, Solar Lentigo, Squamous Cell Carcinoma, Wart.

CHAPTER 3

Elements of Comparison: Learning Task and Simultaneous Presentation

Classification-based learning is one of the most commonly used approaches in category learning paradigms (Markman & Ross, 2003). This approach relies on trials in which a participant is tasked with assigning category membership to a presented stimulus before then receiving feedback on their choice. This is also referred to as the guess-and-correct cycle. There are many benefits for this approach, with a major one being its incorporation of the *testing effect*. The testing effect refers to the idea that learning can be improved through the repeated retrieval of information relative to passive viewing or studying (Roediger & Butler, 2011; Roediger & Karpicke, 2006). While this has often been studied with declarative knowledge, similar benefits may be seen in perceptual learning (Mettler et al., 2019; Thai et al., 2015). By requiring an active judgment on each trial, classification-based learning provides repeated testing opportunities.

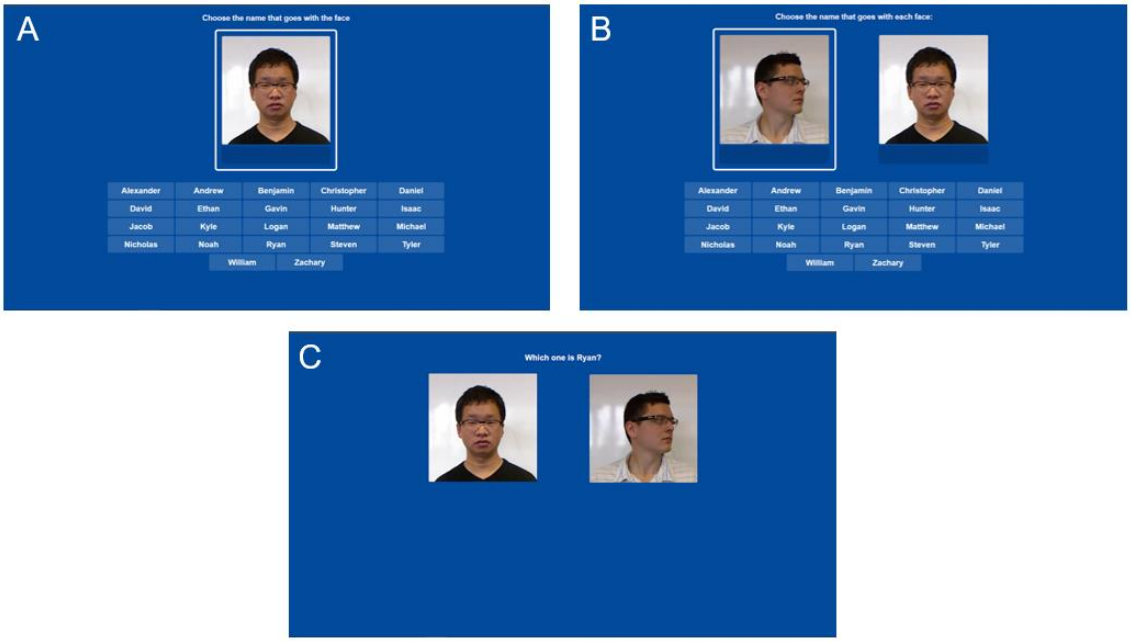
In ordinary experience, the acquisition of categories and the resulting ability to classify new instances seems to occur from exposure to exemplars that occur sequentially, often far apart in time. Accordingly, category learning paradigms conducted in lab-settings also often structure learning in a way that emphasizes only a single item at a time. While comparisons between items may be promoted through these types of sequential paradigms by presenting items within close temporal proximity to one another (e.g., Carvalho & Goldstone, 2015), there are reasons to suspect that the simultaneous presentation of items are especially valuable in enhancing the learning of complex, perceptual classifications.

When items are presented simultaneously, it allows for all information to be available on each trial. This allows the learner to look back and forth between different aspects of the presented stimuli, which, relative to successive, sequential presentations, results in a decreased

memory load. Additionally, if perceptual learning leads to enhanced filtering of information wherein relevant information becomes upweighted and irrelevant information downweighted (cf. Petrov et al., 2005), then the simultaneous presentation of items should allow for easier discovery of invariance (Gibson, 1969). Experimental results have provided evidence that the simultaneous presentation of items can support enhanced differentiation for those items later on (Mundy et al., 2007, 2009) and some work has reported an advantage of learning with simultaneous presentation of category pairs or arrays relative to singular presentations (Andrews et al., 2011; Homa et al., 2014; Jee et al., 2013).

In keeping with the popular classification approach, most studies using active learning methods structure simultaneous learning trials such that label recall is required for one or more presented items (e.g., Andrews et al., 2011; Carvalho & Goldstone, 2014; Homa et al., 2014). However, there are other ways learning trials can be structured to utilize active judgements in a simultaneous presentation that do not emphasize label recall. Instead of selecting category labels for each presented stimulus, one might instead focus on the selection of the category stimulus for a presented category label. In this format, a category label would be presented along with exemplars from multiple categories and participants would be required to select which exemplar matches the presented label. This approach puts a greater emphasis on the comparison and discrimination of presented items, and less on category label recall or recognition. Going forward, I will refer to this trial task as a *paired comparison trial*, and the more traditional classification-based approaches, in which a corresponding category label is selected for each presented stimulus, as simply *single-classification trials* or *dual-classification trials* if two items are presented simultaneously. A visualization of all three trial formats can be seen in Figure 3.

Figure 3
Learning Trial Layouts



Note. The trial layouts used in Experiment 1 and Experiment 2 are shown above. Panel A: Single-Classification Trial ("Choose the name that goes with the face"). Panel B: Dual-Classification Trial ("Choose the name that goes with each face"). Panel C: Paired Comparisons Trial ("Which one is [Category Name]?").

This paired comparison trial format was recently studied in the context of an adaptive learning system. In studies conducted by Jacoby and colleagues (2021), participants received a combination of single item classification and paired comparison trials. Most trials involved single classification trials, but when two errors involving the same pair of categories were made, an adaptively-triggered comparison trial was generated. On these trials, participants were presented with a category label and instructed to choose between two images from the confused categories before resuming standard trials. Ultimately, when compared to a condition containing only single classification trials, the inclusion of paired comparison trials resulted in faster and more efficient learning. While these comparison trials show success when used in conjunction with other classification trials, it raises the question of whether one could effectively learn all

categories in a set through this sort of presentation alone.

One theorized benefit of these paired comparison trials is their ability to give participants the direct opportunity to compare confusable stimuli without the additional cognitive load of considering all other categories in the learning set. Unlike most classification-based learning paradigms, paired comparison trials restrict participants to choosing between a limited set of options (two), rather than considering the entirety of the learning set. This format may enable participants to devote more attention to extracting perceptual patterns and invariants that will advance learning, particularly when learning a large number of perceptual classifications at a time. However, there is also concern that learning in this format alone may be too easy to produce meaningful, long-lasting learning. In particular, it may be difficult to transition from differentiating between only two items to making future classification judgments across all learned categories. Additionally, with guessing rates at 50% for paired comparison trials, there is the potential to induce a misleading sense of fluency, as participants may be able to progress quickly through learning based partially on chance responding.

Although potentially meaningful differences between classification trials and paired comparison trials can be theorized, very little research has directly focused on comparing different active approaches to comparison. The goal of the present chapter was to address this gap in research by comparing classification trial formats to comparison trial formats across two different domains.

Experiment 1

Can the learning of multiple perceptual classifications be achieved through only paired comparison trials—and if so, how might that differ from the more common classification-based learning? In the present experiment, learning and transfer performances were assessed across

three types of learning conditions: a Paired Comparisons condition, a Single-Classification condition, and a Dual-Classification condition. Following learning, participants were tested on their ability to classify previously seen as well as novel images of the learned face categories.

Participants

Undergraduate participants were recruited through the University of California Los Angeles subject pool. We retained and analyzed the data of 75 participants. Participants were awarded partial course credit for their participation.

Stimuli

The stimuli used were five distinct pictures of 22 different human male faces for a total of 110 unique images taken from a larger database (Min et al., 2014). Four images of each of the 22 categories were used in the learning phase. Non–face details such as hairstyle or visible clothing could vary across images within the same category; however, the background and final image size remained identical across all exemplars. The fifth image in each category was set aside for use as a novel stimulus in immediate and delayed posttests. Figure 4 depicts examples of images for one category.

Each learning category consisted of face images from the same person, and each category was identified with a name. The names were chosen to be unremarkable, and were taken from the Social Security list of most common names given in the United States in 2000-2009.

Design & Procedure

Participants completed the study online. Each participant was assigned to one of three learning conditions. All participants completed a learning phase followed by an immediate posttest. Delayed posttests were administered one week later.

Figure 4
Exemplars for a Given Category: "Logan"

Note. Learning images (left) varied in expression, pose, and non-face details. The remaining image (right) was retained for presentation in the posttest assessments

Figure 3 shows an example learning trial from each learning condition. In the Paired Comparisons condition, two faces, each from a separate category, were presented side by side under the prompt "Which one is [Category Name]." Participants were instructed to click on the image of the person named. In the Single-Classification condition, each learning trial contained one face exemplar with all 22 possible name options organized alphabetically below. Participants were required to select the name corresponding to the face presented. In the Dual-Classification condition, two faces were presented side by side, as in the Paired Comparisons condition, but participants chose the name for each face, as in the Single-Classification condition. The faces always belonged to two different categories and participants could choose to classify them in any order.

In all conditions, participants were given up to 40 seconds to answer each learning trial either by selecting an image or name label(s). Immediate feedback was provided after each trial and was given in the form of a green checkmark appearing alongside their answer choice when

correct and a red ‘X’ appearing when incorrect. Additionally, the correct name label was presented under each presented face. Feedback was displayed for up to 10 seconds, however, participants could choose to advance to the next trial at any point.

Categories were interleaved using a blocked-randomized approach, such that each category was presented for classification once per block. In the Paired Comparisons condition, a category was considered to be presented for classification if it was the target of the trial, e.g., if a trial asked “Which one is Benjamin?” then Benjamin was considered to have been presented for classification. The other presented item (the distractor) was chosen randomly from all other 21 possible categories. In all conditions, the order in which categories were presented within each block was randomized across blocks and across participants.

Conditions were equated on time invested in learning. All participants were given 40 minutes of active learning time to progress through the learning phase. Included in these 40 minutes was the time taken to select an answer on each trial, as well as the time spent viewing feedback after each trial. Any other time invested, such as on instruction screens or during periods of inactivity, did not count toward these 40 min.

Immediately following the completion of the learning phase, participants were administered a posttest. One previously seen exemplar per category, as well as one novel exemplar per category, were randomized and presented sequentially for classification. The layout of each test trial was identical to the learning trials in the Single-Classification condition; however, no feedback was given during testing. A delayed posttest, administered one week later, was identical in content and structure to that of the immediate posttest.

Exclusion Criteria

Only data from participants who completed all parts of the experiment (learning phase, immediate posttest, and delayed posttest) were included in the following analyses. Additionally, participants were excluded after data collection if they completed six or more learning blocks at or below chance accuracy. Each learning block consisted of one presentation of each category for identification, resulting in 22 trials per block in the Single-Classification and Paired Comparisons conditions and 11 trials per block in the Dual-Classification condition. In total, four participants were dropped for poor learning performance including one participant assigned to the Single-Classification condition and three assigned to the Paired Comparisons condition.

Dependent Measures and Data Analyses

For each participant, the number of answers given during the learning period was recorded and reported as the total number of classifications invested. Learning trials in the Single-Classification condition and the Paired Comparisons condition required a single response, meaning each trial completed was equivalent to one classification invested; in the Dual-Classification condition, where two separate answers were given on each trial, each learning trial consisted of two classifications.

Posttest accuracies were measured for each participant at the immediate and delayed posttests. Possible interactions between condition and posttest accuracy were tested, as well as possible accuracy differences at both phases of the posttest. Condition comparisons for all measures were compared using ANOVA and other standard parametric statistical methods. Effect sizes are reported for each difference.

Results

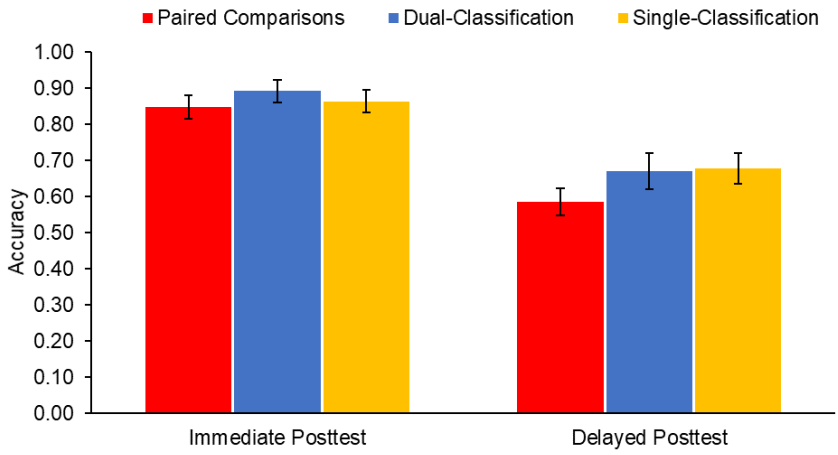
Learning Phase

The number of classifications invested within the 40 minutes of learning time varied across all three conditions. Classifications were highest in the Paired Comparisons condition (M = 798.88, SD = 161.63) followed by the Dual-Classification condition (M = 371.84, SD = 124.91) and lastly the Single-Classification condition (M = 358.16, SD = 162.33). A one-way ANOVA on classifications invested revealed this difference to be reliable, $F(2, 72) = 69.18$, $p < .001$, $\eta_p^2 = 0.66$. Pairwise comparisons revealed that there was a significant difference between the Paired Comparisons condition and the Dual-Classification condition, $t(48) = 10.45$, $p < .001$, $d = 2.95$, as well as between the Paired Comparisons condition and the Single-Classification condition, $t(48) = 9.62$, $p < .001$, $d = 2.72$. There was no reliable difference between the Single- and Dual-Classification conditions, $t(48) = 0.33$, $p = .740$, $d = 0.09$.

Posttest Accuracy

Figure 5 depicts mean accuracy by condition at both phases of the posttest. Accuracy was similar across conditions at both posttest phases. At the immediate posttest, accuracy was highest numerically in the Dual-Classification condition (M = 0.89, SD = 0.15) followed by the Single-Classification condition (M = 0.86, SD = 0.16) and the Paired Comparisons condition (M = 0.85, SD = 0.16). At the delayed posttest, accuracy was highest numerically in the Single-Classification condition (M = 0.68, SD = 0.21), followed by the Dual-Classification condition (M = 0.67, SD = 0.25), and finally the Paired Comparisons condition (M = 0.58, SD = 0.19).

Figure 5
Assessment Accuracy (Experiment 1)



Note. Average assessment accuracy, measured as proportion correct, is graphed for each learning condition at both assessment phases. Error bars indicate +/- 1 standard error of the mean.

A 3 (*condition*) X 2 (*posttest phase*) mixed measures ANOVA was performed on the accuracy data. The results revealed a significant within-subjects effect of posttest phase, such that participants performed better at the immediate posttest than the delayed, $F(1, 72) = 136.94$, $p < .001$, $\eta_p^2 = 0.66$. However, the main effect of condition was not found to be significant, $F(2, 72) = 1.01$, $p = .369$, $\eta_p^2 = .03$, nor was the condition by posttest phase interaction, $F(2, 72) = 1.36$, $p = .263$. $\eta_p^2 = .04$.

Discussion

Participants across all conditions showed significant learning and subsequent generalization of 22 face categories after 40 min of training. Paired comparison trials were completed at a much quicker rate than either classification condition, resulting in over twice as many completed trials in the same amount of time. Regardless of differences during learning, classification accuracy was not found to reliably differ between conditions at either the immediate or delayed posttests. The results from Experiment 1 demonstrate that paired

comparison learning is capable of producing similar learning gains as the more common classification-based approaches, suggesting that participants were not disadvantaged by the lack of recall practice or the restricted answer options on each trial.

Contrary to our initial predictions, we found no benefit of simultaneous presentation, with the Single-Classification condition performing equally as well as both the Dual-Classification and Paired Comparisons learning conditions. One possible explanation for the lack of advantage in the present experiment may be that the categories used were too familiar to the learners. Simultaneous presentations may be maximally beneficial when differences are subtle or difficult to discover, as one of their primary advantages over sequential presentations is the promotion of back-and-forth comparison between items. While the specific faces used in the present study were novel to all participants, given that adult humans are often considered experts in face perception, they may have developed the skill of knowing where to look for the most critical information to determine identity. This may make differences between face categories more easily detectable than they would be in a more novel domain.

A related explanation may be derived from work demonstrating how the value of comparisons can vary as a function of similarity. Specifically, comparison of different items may be maximally beneficial when items are highly similar, and have little to no benefit when presented items are dissimilar (Dwyer & Vladeanu, 2009). Consequently, in the context of category learning, between-category comparisons are less valuable than within-category comparisons when between-category similarity is low (Carvalho & Goldstone, 2014). When it comes to (unaltered) faces, it has been shown that the magnitude of variability that can exist within a face often contributes more to face perceptual errors than do similarities of different faces (Jenkins et al., 2011), suggesting that between-category similarity of unaltered faces may

be relatively low.

While the faces used here had been shown prior to benefit from comparison trials utilizing simultaneous presentation alongside single-item classification trials (Jacoby et al., 2021), those comparisons were adaptively triggered so as to contain the most confusable categories for comparison. (See Chapter 5.) Without intentionally pairing the most similar categories for comparison, between-category comparisons may simply fail to elicit any measurable advantage for this type of stimulus. To follow up on the possibility that the effects of these different learning conditions may interact with characteristics of the to-be learning categories, we tested these learning conditions again in the context of a different domain: skin lesion differentiation.

Experiment 2: Skin Lesion Differentiation

As described in Chapter 2, different types of skin lesions can be visually identified through the integration of multiple specific perceptual features including symmetry, irregularity of border, color, and size. Many of these features can present similarly across different types of lesions, making the differential diagnosis of skin lesions difficult for even very experienced dermatologists. Unlike face identification, most untrained individuals do not have insight as to where to look for the most critical information to determine skin lesion category membership, nor is the most critical information always contained in the same locations across different categories. When compared to the facial identity categories in Experiment 1, skin lesion categories are characterized by much greater between-category similarity.

Experiment 2 had two primary goals. The first was to test the robustness and efficacy of paired comparison learning in a new, different domain. The second goal was to investigate

whether using categories with higher between-category similarity, would reveal any benefit to simultaneous between-category presentations that was not found in Experiment 1.

Method

Participants

Undergraduate participants were recruited through the University of California, Los Angeles subject pool. We retained and analyzed the data from 90 participants. Participants completed the experiment online and were awarded partial course credit for their participation.

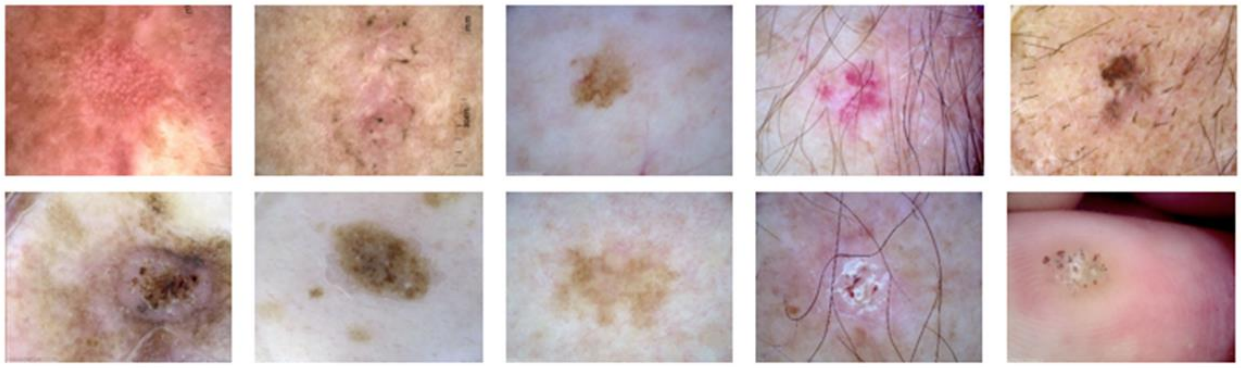
Stimuli

Figure 6 contains example stimuli from each category. Stimuli consisted of dermoscopic images of 10 different skin lesion categories including four cancerous skin lesion categories (Basal Cell Carcinoma, Lentigo Maligna Melanoma, Nodular Melanoma, Squamous Cell Carcinoma) and six benign categories (Actinic Keratosis, Benign Nevus, Haemangioma, Solar Lentigo, Seborrheic Keratosis, Wart). All images were obtained from the MoleMap Database The number of instances available to appear in learning varied per category with each category containing anywhere from 19 to 165 unique images, for a total of 715 items available in learning. An additional two exemplars per category were set aside to use as novel items at the posttests.

Design & Procedure

To ensure that participants did not enter the experiment with significant knowledge of the to-be-learned skin lesion categories, all participants were administered a pretest before beginning the learning phase. Consistent with the posttest format, two items per category were presented sequentially for classification. Presentation order was randomized and no feedback was provided.

Figure 6
Skin Lesion Categories



Note. One example from each of the ten categories are shown above. From left to right, top to bottom: Actinic Keratosis, Basal Cell Carcinoma, Benign Nevus, Haemangioma, Lentigo Maligna Melanoma, Nodular Melanoma, Seborrheic Keratosis, Solar Lentigo, Squamous Cell Carcinoma, Wart.

Participants were then assigned to one of three possible learning conditions: the Paired Comparisons condition, Dual-Classification condition, or Single-Classification condition. The structure of learning trials and the learning phase procedure was identical to that of Experiment 1. After 40 minutes of active learning, participants were administered an immediate posttest containing 2 novel exemplars per category for a total of 20 items. An identical posttest was administered one week later.

Exclusion Criteria

Only data from participants who completed all parts of the experiment (pretest, learning phase, immediate posttest, and delayed posttest) were included in the following analyses. Participants who scored 30% or greater on the pretest were disqualified and did not participate in the rest of the experiment.

Participants were excluded after data collection if they completed 10 or more learning blocks with chance or below chance accuracy. Eight participants were dropped for poor learning

performance, including 5 participants assigned to the Paired Comparisons condition, 2 participants assigned to the Dual-Classification condition, and 1 participant assigned to the Single-Classification condition.

Dependent Measures and Data Analyses

For each participant, the number of classifications invested in learning was recorded, as well as accuracies on the pretest, immediate posttest, and delayed posttest. Possible interactions between condition and posttest accuracy were tested, as well as accuracy differences at both phases of the posttest. Condition comparisons for all measures were compared using ANOVA and other standard parametric statistical methods.

Results

Learning Measures

Classifications Invested

The number of classifications invested within the 40 minutes of learning time varied across all three conditions. Those in the Paired Comparison condition invested the most (M = 553.53, SD = 190.46) followed by the Dual-Classification condition (M = 396.53, SD = 139.30) and the Single-Classification Classification condition (M = 369.33, SD = 139.14). A one-way ANOVA confirmed a significant difference between conditions, F (2, 87) = 11.86, $p < .001$, η_p^2 = 0.21 and subsequent contrasts determined the difference between the Paired Comparisons and Dual-Classification conditions to be significant, $t(58) = 3.64$, $p = .001$, $d = 0.94$, as well as between the Paired Comparisons and Single-Classification condition, $t(58) = 4.28$, $p < .001$, d = 1.10. The difference between the Dual-Classification and Single-Classification conditions was not reliable, $t(58) = 0.76$, $p = .452$, $d = 0.20$.

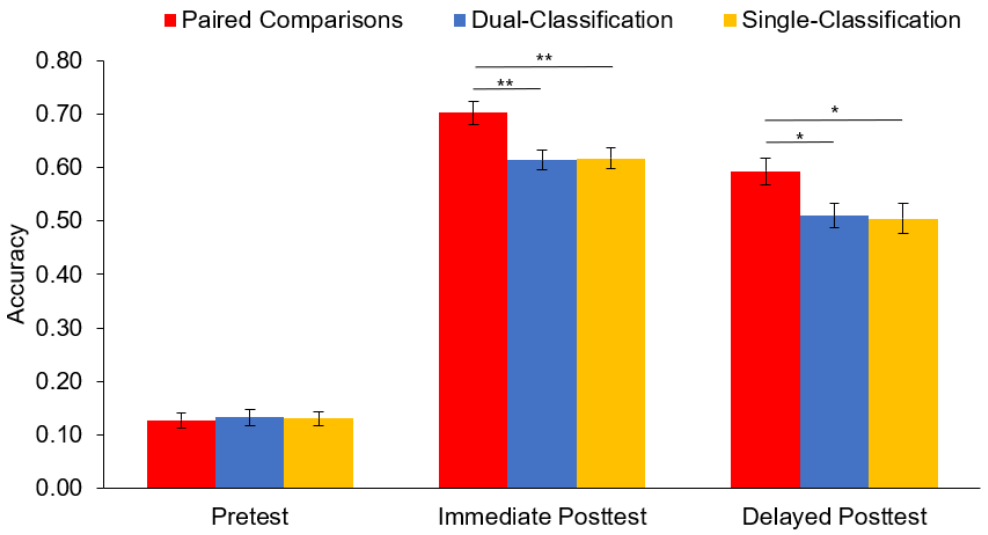
Assessment Accuracy

Due to a technical error, 13 participants in the Single-Classification condition and 13 participants in the Dual-Classification condition were unintentionally exposed to a Lentigo Maligna Melanoma assessment item once during the learning phase, making the item no longer novel at posttests. For these 26 participants, we recalculated their assessment accuracy using only the remaining 19 items. Analyses were conducted on both the adjusted and unadjusted scores. There was no difference in overall significance between the adjusted and unadjusted scores for either the omnibus ANOVA or the subsequent paired contrasts at either phase of the posttest, and differences in effect size were minor. Reported here are the adjusted assessment scores. Analyses with unadjusted scores can be found in Appendix A.

Figure 7 depicts average accuracy by condition at the pretest, immediate posttest, and delayed posttest. Pretest scores did not differ across conditions, with an average pretest score of 13% in all conditions, $F(2, 87) = 0.03$, $p = .967$. At the immediate posttest, accuracy was greatest in the Paired Comparisons condition ($M = 0.70$, $SD = 0.12$) relative to the Dual-Classification ($M = 0.61$, $SD = 0.10$) and Single-Classification ($M = 0.62$, $SD = 0.11$) conditions. At the delayed posttest, accuracy was highest in the Paired Comparisons condition ($M = 0.59$, $SD = 0.13$), followed by Dual-Classification condition ($M = 0.51$, $SD = 0.13$) and the Single-Classification condition ($M = 0.50$, $SD = 0.15$).

A 3 (*condition*) X 2 (*posttest phase*) mixed measures ANOVA was performed on the posttest accuracy scores. There was a reliable within-subjects effect of test phase, such that participants performed better on the immediate posttest than the delayed, $F(1, 87) = 50.39$, $p < .001$, $\eta_p^2 = 0.37$, as well as a reliable between-subjects effect of learning condition, $F(2, 87) = 7.11$, $p = .001$, $\eta_p^2 = .14$. We found no reliable evidence of an interaction between posttest phase and condition, $F(2, 87) = 0.02$, $p = .976$.

Figure 7
Assessment Accuracy (Experiment 2)



Note. Average assessment accuracy, measured as proportion correct, is graphed for each learning condition at each assessment phase. Error bars indicate +/- 1 standard error of the mean. Statistical significance is indicated with asterisks: * $p < .05$, ** $p < .01$, ***$p < .001$.

Planned pairwise contrasts between learning conditions were conducted at each posttest phase. At the immediate posttest, an independent samples t-test revealed that the accuracy advantage of the Paired Comparisons condition was significant relative to both the Dual-Classification condition, $t(58) = 3.01$, $p = .004$, $d = 0.79$**,** and the Single-Classification conditions, $t(58) = 2.92$, $p = .005$, $d = 0.73$. There was not a reliable difference between the Dual-Classification and Single-Classification conditions, $t(58) = -0.08$, $p = .939$, $d = 0.09$. This pattern was consistent at the delayed posttest, with a significant difference found between the Paired Comparisons condition and the Dual-Classification condition, $t(58) = 2.47$, $p = .016$, $d = 0.63$, as well as between the Paired Comparisons condition and Single-Classification condition, $t(58) = 2.39$, $p = .020$, $d = 0.64$. There was no reliable difference between the Dual-Classification and Single-Classification conditions, $t(58) = 0.17$, $p = .866$, $d = 0.07$.

Discussion

Experiment 2 tested Paired Comparisons learning, Dual-Classification learning, and Single-Classification learning in the context of skin lesion identification. Our primary goal was to demonstrate whether paired comparisons could facilitate the learning of a large number of classifications in an entirely different domain. Our secondary goal was to evaluate if a simultaneous-presentation advantage could be revealed when using more difficult stimuli in a domain novel to all learners.

Consistent with Experiment 1, when given the same amount of time to learn, those who learned with paired comparison trials completed more learning classifications than the Dual-Classification or Single-Classification conditions. However, unlike Experiment 1 which showed equal performance on subsequent assessments, here, accuracy on both the immediate and delayed posttests reliably favored the Paired Comparisons learning condition over either classification-based condition (all medium effect sizes). When it comes to evaluating whether there was an effect of simultaneous presentation, while we do see an advantage of Paired Comparison learning, which utilizes a simultaneous presentation, we also found that there was no advantage of dual-item presentation over single-item presentation in the classification conditions. This suggests that it is the task of the Paired Comparisons learning rather than the simultaneous nature of the presentation that is primarily responsible for the observed performance difference.

General Discussion

Across two experiments, we tested different approaches to learning and comparison in a category learning task containing naturalistic categories. In Experiment 1, the learning of 22 facial-identification categories was facilitated equally well through a paired comparisons

approach that emphasized the discrimination of two different exemplars as the more standard single-item classification approach and a dual-item classification approach. In Experiment 2, an advantage of paired comparisons learning was revealed when the learning domain was changed to a 10-category skin lesion differential diagnosis task.

Why did we observe a benefit in the second experiment that was not present in the first? By using a domain characterized by greater similarity between categories and ensuring that participants in Experiment 2 were not already perceptual experts with the material, it may have increased the difficulty of the task and consequently increased the value of directly comparing items together. While participants in both experiments were able to learn to significantly above chance performance, the average learning accuracy dropped considerably from Experiment 1 to Experiment 2, even despite the smaller number of categories learned (Exp. 1: Paired Comparisons $M = 0.93$, Dual-Classification $M = 0.60$, Single-Classification $M = 0.53$; Exp. 2: Paired Comparisons, $M = 0.82$, Dual-Classification, $M = 0.38$, Single-Classification, $M = 0.40$). These differences confirm that skin lesion differentiation was a more difficult task than facial identification.

If participants were taking advantage of the opportunity to compare items more when the task was difficult, then we might see that reflected in the number of trials they invested within the 40 min training period. When comparing the number of classifications invested in the face-identification training versus the skin lesion differentiation training, we see that the overall pattern of results stayed consistent between experiments with participants in the Paired Comparisons condition completing significantly more trials than in either classification condition; however, the magnitude of this effect diminishes considerably between experiments with the effect size dropping by 68% (Exp. 1 $\eta_p^2 = 0.66$, Exp. 2 $\eta_p^2 = 0.21$). Notably, this

difference in effect appears to be primarily driven by changes in the number of trials invested in the Paired Comparisons condition specifically, with participants completing 30% fewer comparison trials in Experiment 2 than in Experiment 1, compared to a 6% increase in classifications invested in the Dual-Classification condition and a 3% increase in the Single-Item condition. This supports our hypothesis that the paired comparisons benefit in Experiment 2 arose from participants engaging more deeply with comparisons of skin lesions than of faces.

Interestingly, despite the opportunity to engage in direct comparison between presented items, the time spent on each dual-classification trial remained roughly consistent regardless of domain. Accordingly, learning through the dual-classification approach did not yield any performance gains relative to the single-item approach, even when the stimuli were made more difficult. While this is inconsistent with prior research demonstrating an advantage of simultaneous presentation (e.g., Andrews et al., 2011), it should be noted that other work has reported only "marginally" significant or non-reliable advantages (Carvalho & Goldstone, 2014; Higgins & Ross, 2011, Kang & Pashler, 2012), suggesting that if a simultaneous presentation benefit exists, it may be small or dependent upon other factors in learning.

Importantly, advantages of simultaneous presentation are only likely to exist if a participant engages in a comparison between the presented items. While participants were able to compare the two items on a Dual-Classification trial, they were not explicitly instructed to do so. Given the nature of the task, each classification could be made without consideration to the other stimulus presented. Learners are often considered to be conservative with regard to their cognitive effort, often only learning and engaging with the minimum amount of information necessary to complete a task (Payne et al., 1993). Given that engaging in a meaningful comparison of items was not a necessary aspect of these learning trials, it is probable that the

reason we do not see an advantage is that participants did not compare the items presented to them.

Recent work by Patterson and Kurtz (2020) provides a complementary explanation. In a study focused on relational category learning, they demonstrated that when items were presented passively (i.e. with the appropriate category labels displayed alongside each stimulus) that there was an advantage of simultaneous presentation over sequential; however, when presented items required a classification judgment, the simultaneous presentation advantage disappeared. The authors assert that this active, classification approach put too much emphasis on correctly identifying each item that it diluted the opportunity to compare and ultimately undermined any potential advantage of the simultaneous presentation.

Here, the paired comparison trial format may have avoided shortcomings of combining active learning trials with opportunities to compare by requiring participants to engage in comparison as a requirement of the task, as opposed to an optional opportunity. Further, by relieving the need for label recall, participants may devote more cognitive resources to comparing the items, while maintaining the opportunity to see each presented exemplar labeled in the feedback.

The outcome of paired comparisons learning may be particularly impressive when considering the format of the assessments. While participants in the classification conditions received practice identifying the appropriate label throughout the entire learning phase, participants in the Paired Comparisons condition had no practice with this format until they reached the assessment itself; thus demonstrating an impressive ability to generalize what they learned across tasks.

There are a few limitations and future directions that this work inspires. In the present

study, paired comparison trials were shown to be quicker to complete than classification-based trials; however, we are limited here only to classification conditions in which all possible category labels were available for selection. We cannot say whether paired comparison trials were quicker to answer due to the nature of the task (selecting between presented images to match a label rather than selecting among labels to match an image) or due to the limited number of options necessary to consider. If participants in the classification conditions were required to select a label from a limited selection of options rather than the entire set, it is possible that we could see a greater number of trials completed due to the reduced cognitive load of no longer considering additional categories on each trial or, possibly, due to less time needed to locate the desired answer choice among the list of possible labels.

Even greater than the condition differences in the number of answers given in learning is the difference in the number of images learners saw in each condition. Not only were paired comparison trials completed more quickly, but they also provided exposure to (and subsequent feedback on) two separate category instances for each classification invested. In order for participants in the classification conditions to be exposed to the same number of images in learning, they would need to complete trials twice as quickly as those in the Paired Comparisons condition.

Prior research has shown that exposure to a greater number of category instances in learning can improve one's ability to classify novel instances in the future (Homa et al., 1991, 2008). As a consequence of controlling for time, as opposed to the number of items exposed to, participants who completed trials more quickly were able to see more skin lesion images. As a result, while participants in the classification conditions in Experiment 2 were exposed to 337 unique category exemplars on average, those in the Paired Comparisons condition were exposed

to 531. It is possible that part of the advantage of paired comparisons learning is derived not just from the value of comparison, but also from a design that allows for more training instances to be presented within the same amount of time. That said, posttest performance was not found to correlate with the number of unique exemplars seen in learning within any of the conditions (all $p > .10$), suggesting that if seeing more unique instances led to a performance benefit in the present work, its effect is likely small. Nonetheless, future work should aim to compare these learning approaches under conditions that equate on learning investments other than time to further isolate the driving factor(s) behind the paired comparison learning benefit.

Conclusion

Training with paired comparison trials is an effective way to promote the learning of multi-category classifications in complex, naturalistic domains. While simultaneous presentation alone is not enough to induce a learning benefit over the traditional single-item classification approach, learning through paired comparisons requires learners to engage with both items presented to them and may be especially advantageous in the learning of difficult to differentiate categories, such as in the classification of cancerous and benign skin lesions.

CHAPTER 4

Dissection of Paired Comparison Learning

Given the success of paired comparison learning in Experiment 2, as well as the general novelty of the learning format for learning large numbers of categories, the goal of this chapter was to more closely examine how learning progresses through the paired comparison format, particularly in the context of skin lesion differentiation. In Experiment 3, I tested how learning advances for each presented category on a trial-by-trial basis by breaking down each trial into its separate "target" and "distractor" components. In Experiment 4, I aimed to explain how performance improvements in paired comparison learning were achieved by measuring how category representations change in response to training.

Experiment 3

The results of Experiments 1 and 2 provide insight into how final performance may differ between different learning and comparison trial types; however, it does little to elucidate how learning may differ on a trial-to-trial basis. We can start by considering what information might be learned on each of the studied learning trial formats. Most obviously, on a single-classification trial, learning will advance primarily for whichever category is presented on that trial. In a dual-classification approach, the two categories presented will both benefit from presentation. Further, one can also reasonably assume that the learning strength for each of the two presented categories will increase a similar amount per trial as both categories will benefit from equivalent retrieval practice and subsequent corrective feedback. In other words, it is unlikely that the participant is learning more about one presented item than the other (assuming similar learning strengths at the beginning of the trial). This is supported by the results of Experiments 1 and 2 in which it was demonstrated that the dual-classification approach yielded a

very similar number of classifications in learning as the single-classification approach while maintaining the same posttest performance.

It is less clear what is learned about each presented category on a paired comparison trial. Like the dual-classification approach, the paired comparison format provides learners the opportunity to see two different category examples, while also providing critical feedback regarding each item's category membership; however, given the prompt of each trial, only one category is framed as the target, with the other presented category fulfilling the role of a distractor. A breakdown of the trial layout can be seen in Figure 8. Although the opportunity to learn about both presented categories is equal, it is possible that the different framing would lead to disparities in the extent to which learning progresses for each presented category.

Figure 8
A Paired Comparison Learning Trial With Feedback

Note. This trial asks: "Which one is Solar Lentigo?", resulting in Solar Lentigo occupying the role of "target". The left image depicts Solar Lentigo (target) and the right image depicts Lentigo Maligna Melanoma (distractor). Feedback is presented identically for both images.

There are a few reasons for why differences in the processing of targets and distractors in this format may be expected. As mentioned in Chapter 3, learners often only engage with the

minimum amount of information necessary to complete a task (Payne et al., 1993). Early in learning, when the learning strength for all categories is low, items are likely compared closely to complete each trial–which should lead to benefits in the perception of both categories. However, as learning progresses and close comparison of items becomes less necessary, attention may be devoted primarily to the target category. This may suggest that while the target category will benefit from engaging in an active learning process (matching the label to an image), the learning of the distractor may be much more passive and rely primarily upon viewing the labeled image in feedback.

Additionally, given that knowing the category label of the distractor image is not necessary for correctly completing the trial (i.e., one need not know what the distractor category is, just that it is not the target category), participants may be less inclined to devote attention to that category during feedback. Finally, while both presented categories are displayed alongside their category label after each trial, the target of a trial gets the benefit of having its category label visually presented twice (once in the instructions and once in feedback). If a category were to regularly appear as a target, rather than a distractor, the increased number of times a learner is exposed to its label may give the impression that that category is more important or possibly more common than the distractor, which in turn could result in an overestimation of likelihood that that category is presented in future encounters.

In the present experiment, we investigated how learning advances in paired comparison training, and in particular, whether the specific role of a category on the trial (target vs distractor) differentially affects the learning of that category. To evaluate this, participants were tasked with learning the perceptual classification of ten dermatological lesion categories while the frequency with which a specific category showed up in learning as a target or distractor was manipulated

across three different learning conditions.

In the Always-Never learning condition, half of the to-be learned categories were assigned to appear in learning only as targets, whereas the other half appeared only as distractors. In the Often-Rarely learning condition, half of the categories showed up as targets on 75% of trials and as distractors on 25% of trials, with the remaining categories following the opposite scheduling. Finally, in the Equal Split condition, all categories showed up in learning equally as often as the target of the learning trial and as the distractor. Within each condition, we compared the performance on categories prioritized as targets to the performance on categories prioritized as distractors. If the design of these trials provides an asymmetric learning gain that benefits target categories relative to distractor categories, then we would expect to see a positive relationship between the frequency of presentations as a target and final assessment performance. Further, we expected to see the largest disparity in performance among categories in the Always-Never condition, followed by a smaller difference in the Often-Rarely condition, and no difference in the Equal Split condition.

Method

Participants

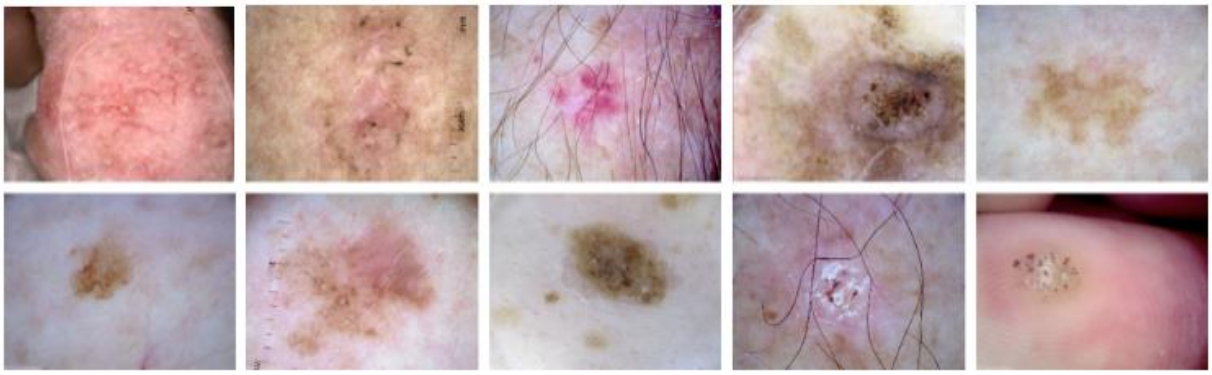
104 undergraduate psychology students from the University of California Los Angeles completed this experiment. Participants had no particular medical background or training and received partial course credit for their participation

Materials

Stimuli consisted of dermoscopic images of 10 different skin lesion categories including four cancerous skin lesion categories and six benign categories. Four exemplars per category were set aside as novel stimuli to be used in assessments.

Categories were divided into two separate lists, each containing five categories. List 1 contained: Actinic Keratosis, Basal Cell Carcinoma, Haemangioma, Nodular Melanoma, and Solar Lentigo. List 2 contained Benign Nevus, Lentigo Maligna Melanoma, Seborrheic Keratosis, Squamous Cell Carcinoma, and Wart. The division was made such that each list contained two cancerous categories (one form of melanoma and one type of carcinoma), three benign categories, one type of keratosis, and one type of lentigo. Exemplars from each list are shown in Figure 9. All participants learned both lists of categories.

Figure 9
Example Category Images by List Assignment



Note. The top row depicts List 1 categories in alphabetical order from left to right; the bottom row depicts List 2 categories.

Design & Procedure

The experiment used a mixed measures design, with three between-participant learning conditions and two within-participant priority lists. Each participant was assigned to one of three learning conditions: Always-Never Learning, Often-Rarely Learning, or Equal Split Learning. Within each learning condition one list of categories was designated as the Target-Priority list and the other was designated as the Distractor-Priority list. This assignment was counterbalanced within each learning condition to control for possible differences in list difficulty.

Learning Phase

All participants completed a series of paired comparison trials, in which an exemplar from two separate categories would be presented side by side and participants would be asked to identify a given target through a prompt asking "Which is [Category Name]?" An example of this trial can be seen in Figure 8.

In the Always-Never learning condition, when any of the five categories designated as the Target-Priority list appeared on a trial they would always be the target and never be the distractor. Conversely, the five categories designated as the Distractor-Priority list would always appear as the distractor but never the target. As a concrete example, if solar lentigo was on the list of categories designated as the Target-Priority list and lentigo maligna melanoma was on the list of categories designated as the Distractor-Priority list, a participant could receive a trial that displayed one solar lentigo image and one lentigo maligna melanoma image side by side and asks "Which is Solar Lentigo?" but never a trial that presents the same images and asks "Which is Lentigo Maligna Melanoma?"

In the Often-Rarely learning condition, categories designated as the Target-Priority list would appear as the target of the trial on 75% of presentations and as the distractor on 25%; the reverse was true for categories designated as the Distractor-Priority list.

Finally, in the Equal Split learning condition, all appeared equally often as target or as the distractor. If solar lentigo and lentigo maligna melanoma were presented together in the trial, it was equally likely that the prompt would be "Which is Solar Lentigo" or "Which is Lentigo Maligna Melanoma?". Although there is no difference between the Target-Priority and Distractor-Priority lists in the Equal Split condition, category lists were still designed as separate priority lists to allow for subsequent analyses.

Participants in all conditions completed 400 learning trials containing a total of 800 images. Importantly, all categories were shown 80 times in learning regardless of the learning condition or target/distractor-priority list. Every trial contained one category from the Target-Priority list and one category from the Distractor-Priority list. Categories were presented in a randomized order. Participants were informed that they would be learning 10 different categories of skin lesions, but they were not made aware of how often a category was to be shown as the target or distractor.

Feedback was provided immediately after each trial and indicated whether the answer was correct/incorrect, as well as labeled both presented images. Participants were given 40 seconds to complete each learning trial and up to 10 seconds to view feedback.

Testing Phase

Participants completed an assessment at three different timepoints: before learning, immediately following learning, and after a one-week delay. Assessments required the classification of individually presented skin lesion images, where on each test trial, one image was shown with all ten category labels organized alphabetically below. Participants had 40s to select a category label on each trial; no feedback was given. Each test contained four items from each category, and only novel (previously unseen) exemplars were used.

Exclusion Criteria

To ensure that participants did not have considerable knowledge of these skin lesion categories prior to starting the experiment, those who scored 30% or greater on the pretest were disqualified and did not participate in the rest of the experiment. Only data from participants who completed all parts of the experiment (pretest, learning phase, immediate posttest, and delayed posttest) were included in the following analyses.

Participants were excluded after data collection if they failed to achieve an average accuracy in the learning phase of 65% (chance accuracy: 50%). In total, 14 participants were excluded for poor learning performance (Always-Never condition: n = 2; Often-Rarely condition: n = 5; Equal Split condition: n = 7). Data from 90 participants were retained.

Dependent Measures and Data Analyses

Accuracy on learning trials and assessments was recorded for all participants and compared across learning conditions. Performance on the posttests was also divided between categories designated as the Target-Priority list and the Distractor-Priority list. Differences between priority lists were compared within each learning condition. To determine whether observed differences in accuracy were due to changes in perception rather than response behavior, we also calculated and compared false alarm rates and sensitivity (d'). This allowed us to assess if guessing behavior was biased toward categories on either priority list. All analyses were conducted using standard parametric measures. Effect sizes are reported for each difference.

Results

Learning Accuracy

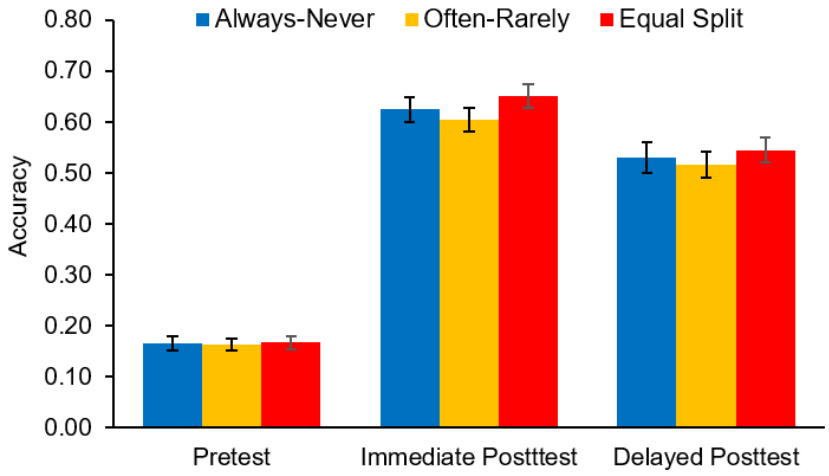
All participants completed 400 paired comparison trials in learning. Average accuracy during the learning phase was highest in the Always-Never condition (M = 0.81, SD = 0.06) followed by the Often-Rarely condition (M = 0.80, SD = .07) and the Equal Split condition (M = 0.78, SD = .07). A one-way ANOVA determined that differences in learning accuracy were not reliable, $F(2, 87) = 1.78$, $p = .175$, $\eta_p^2 = 0.04$.

Assessment Performance

Overall Accuracy

Figure 10 depicts average accuracy across all assessment items at the pretest, immediate posttest, and delayed posttest for each learning condition. Average performance on the pretest assessment was 16.44% (SD = 6.77), and did not differ between conditions, $F(2, 89) = 0.03$, $p = .972$, $\eta_p^2 = .001$.

Figure 10
Overall Assessment Accuracy (Experiment 3)



Note. Accuracy, measured as proportion correct, for each of the three learning conditions at each assessment phase. Error bars indicate +/- 1 standard error of the mean.

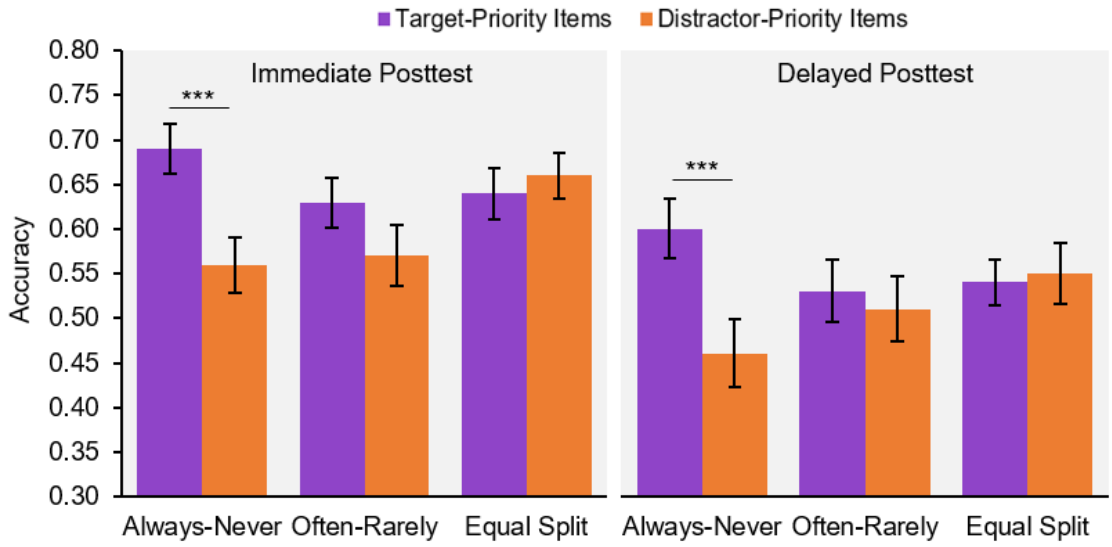
Posttest accuracy was similar across all three conditions at the immediate posttest (Equal Split: $M = .65$, SD = .13; Always-Never: $M = .62$, $SD = .13$; Often-Rarely: $M = .60$, $SD = .13$) and at the delayed posttest (Equal Split: $M = .54$, $SD = .13$; Always-Never: $M = .53$, $SD = .17$; Often-Rarely: $M = .52$, SD = .15). A 3 (learning condition) X 2 (posttest phase) mixed measures ANOVA was conducted on the assessment scores. There was a significant main effect of posttest phase, such that participants performed worse on the one-week delayed posttest than on the immediate posttest regardless of learning condition, $F(1, 87) = 54.37$, $p < .001$, $\eta_p^2 = 0.39$. There

was not reliable main effect of the learning condition, $F(2, 87) = 0.67$, $p = .513$, $\eta_p^2 = 0.02$, nor a significant learning condition by posttest phase interaction, $F(2, 87) = 0.17$, $p = .844$, $\eta_p^2 = 0.004$.

Target vs. Distractor Accuracy

Posttest items were then divided into the categories that had shown up in learning as Target-Priority items and Distractor-Priority items. Figure 11 shows average accuracy on Target-Priority and Distractor-Priority categories for each learning condition at both post-learning assessments.

Figure 11
Posttest Accuracy for Targets vs. Distractors



Note. Accuracy (proportion correct) for categories prioritized as targets and for categories prioritized as distractors for each learning condition. Error bars indicate +/- 1 standard error of the mean. Statistical significance is indicated with asterisks: * $p < .05$, ** $p < .01$, ***p < .001.

A 2 (Priority List) by 2 (Posttest Phase) by 3 (Learning Condition) mixed measures ANOVA was conducted on assessment scores. Analyses revealed a significant main effect of posttest phase, such that all items were classified with lower accuracy at the delayed posttest

relative to the immediate posttest, regardless of learning condition, $F(1,87) = 54.37$, $p < .001$, $\eta_p^2 = 0.39$. Posttest phase did not interact with learning condition, $F(2, 87) = 0.17$, $p = .844$, $\eta_p^2 = 0.004$, or with priority list, $F(2, 87) = 0.001$, $p = .980$, $\eta_p^2 = 0$. Additionally, no three-way interaction was found, $F(2, 87) = .0.52$, $p = .599$, $\eta_p^2 = 0.01$. A significant interaction was found between priority list and learning condition, $F(2, 87) = 4.56$, $p = .013$, $\eta_p^2 = 0.10$. To further evaluate this relationship, the effect of priority list was assessed separately for each learning condition.

In the Always-Never condition, categories that were assigned to the Target-Priority list were more accurately classified than the categories assigned to the Distractor-Priority list at the immediate posttest (Target-Priority: $M = 0.69$, $SD = 0.15$, Distractor-Priority: $M = 0.56$, $SD = 0.17$), as well as at the delayed posttest (Target-Priority: $M = 0.60$, $SD = 0.19$, Distractor-Priority: $M = 0.46$, $SD = 0.21$). Analyses confirmed the average difference between priority lists to be significant with a large effect size, $t(29) = 4.40$, $p < .001$, $d_z = 0.80$.

In the Often-Rarely condition, accuracy at the immediate posttest favored the categories prioritized as targets ($M = 0.63$, $SD = 0.15$) over those prioritized as distractors ($M = 0.57$, $SD = 0.19$). The same pattern was found at the delayed posttest (Target-Priority: $M = 0.53$, $SD = 0.19$; Distractor-Priority: $M = 0.51$, $SD = 0.20$). However, analyses revealed that these differences between priority lists were not reliable, $t(29) = 0.94$, $p = .357$, $d_z = 0.17$.

Finally, in the Equal Split learning condition, performance on categories designated as the Target-Priority list was expectedly similar to those designated as the Distractor-Priority list at both the immediate posttest (Target-Priority: $M = 0.64$, $SD = 0.16$; Distractor-Priority: $M = 0.66$, $SD = 0.14$) and delayed posttest (Target-Priority: $M = 0.54$, $SD = .14$; Distractor-Priority: $M = 0.55$, $SD = 0.19$). No reliable effect of priority list was found, $t(29) = -0.38$, $p = .705$, $d_z = 0.07$.

Figure 12

Difference in Accuracy Between Target-Priority and Distractor–Priority lists by Priority List Probability Difference

Note. Group mean difference in accuracy between Target-Priority and Distractor–Priority lists is graphed by Priority List Probability Difference for both the immediate posttest and delayed posttest.

We also looked at the priority list by learning condition interaction in a different way, to gain insight into the trend across conditions. These results are shown in Figure 12, which displays group means for target-priority categories vs. distractor-priority categories by condition for both immediate and delayed posttest target-accuracy differences. Individual subject data varied greatly, but a single linear fit accounted well for the group mean data, $F(1, 4) = 58.56$, $p = .002$, with an R^2 value of .94. The y-intercept for the combined data was -.018, very close to the theoretical value of no difference for target list vs. distractor list in the Equal Split condition. Even with such a small number of data points, the slope of .138 was highly reliably different from zero, $t(4) = 7.65$, $p = .0015$, indicating that as the target-frequency disparity between target-

priority and distractor-priority categories increased, the performance difference for the target and distractor lists increased linearly. Separate linear fits for the immediate and delayed posttest data yielded remarkably similar parameter estimates (slope .142 vs. .135; y intercept -.019 vs. -.016), suggesting that, despite overall accuracy decrements from immediate to delayed tests, the difference between target and distractor list performance showed an approximately invariant effect in the two assessments.

False Alarm Rates

In this multi-category classification paradigm, misclassified trials were recorded as a "miss" for the presented category, a "false alarm" for the category whose label was incorrectly selected, and a "correct rejection" for all remaining 8 categories. A false alarm rate was calculated for each category and, consistent with the posttest accuracy analyses, was averaged across the list of categories prioritized as targets in learning and the list of categories prioritized as distractors. This enabled us to test whether superior performance on Target-Priority categories might have resulted from participants, when unsure, guessing those categories more often than Distractor-Priority categories.

A 2 (Priority List) by 2 (Posttest Phase) by 3 (Learning Condition) mixed measures ANOVA was conducted on false alarm rates. Consistent with the decrease in overall accuracy from immediate to delayed test, there was a reliable main effect of posttest phase, with more false alarms committed at delayed posttest than at immediate posttest, $F(1, 87) = 171.22$, $p < .001$, $\eta_p^2 = 0.66$. There was no main effect of priority list ($p = .808$) or learning condition ($p = .579$), and no reliable 2-way or 3-way interactions (all $p > .20$), suggesting that whether a category more frequently occupied a target or distractor position had little or no effect on subsequent classification response biases.

Sensitivity (d')

Accuracy (hit rates) and false alarm rates were used to calculate sensitivity, measured as d', for each category before being averaged across priority lists. A 2 (Priority List) by 2 (Posttest Phase) by 3 (Learning Condition) mixed measures ANOVA was conducted on d' scores. There was a main effect of posttest phase, $F(1, 87) = 80.66$, $p < .001$, $\eta_p^2 = 0.48$, indicating greater sensitivity at immediate posttest. There was a main effect of priority list, $F(1, 87) = 4.38$, $p = .039$, $\eta_p^2 = 0.05$, such that d' was higher for items prioritized as targets rather than distractors. The priority list by learning condition interaction was found to be marginally significant, $F(2, 87) = 2.68$, $p = .074$, $\eta_p^2 = 0.06$. All other two-way and three-way interactions were not reliable, all $p > .400$.

As expected given the observed differences in the accuracy and the lack of differences in false alarm rates, the general pattern of results for d' paralleled the accuracy results. Within the Always-Never condition, d' was significantly greater for categories prioritized as the target (immediate posttest: $M = 2.29$, $SD = .60$; delayed posttest: $M = 1.84$, $SD = .70$) than categories prioritized as the distractor (immediate: $M = 1.98$, $SD = .67$; delayed: $M = 1.51$, $SD = .73$), $t(29) = 3.31$, $p = .002$, $d_z = 0.60$. In the Often-Rarely condition, d' was greater for Target-Priority categories (immediate: $M = 2.12$, $SD = .57$; delayed: $M = 1.65$, $SD = .66$) than Distractor-Priority categories (immediate: $M = 1.94$, $SD = .73$; delayed: $M = 1.63$, $SD = .72$), though this difference was not found to be reliable, $t(29) = 1.29$, $p = .207$, $d_z = 0.15$. Finally, in the Equal Split condition differences between Target-Priority list (immediate: $M = 2.21$ $SD = .61$; delayed: $M = 1.71$, $SD = .50$) and Distractor-Priority lists (immediate: $M = 2.26$, $SD = .60$; delayed: 1.72, $SD = .68$) did not differ, $t(29) = -0.30$, $p = .768$, $d_z = 0.05$.

Discussion

The present study investigated whether the role of categories in a paired comparison task influences learning. In all three learning conditions, Always-Never learning, Often-Rarely learning, and Equal Split learning, novice participants improved their ability to classify difficult skin lesion images after a short learning period. Notably, all learning conditions produced roughly equivalent learning gains.

Despite similar overall performance, there were clear differences between conditions in how overall accuracy was achieved. The Always-Never condition showed significantly greater learning gains, with a large effect size, for categories that appeared as targets than for those only shown as distractors, whereas no reliable differences were observed between targets and distractors in the Often-Rarely and Equal Split conditions. The numerical difference, t-value, and effect size in the Often-Rarely condition fell between those of the Always-Never and Equal Split conditions.

Notably, when any image appeared in any comparison trial, there was no information in the display that indicated its role as target or distractor. Only with feedback was the role revealed; even then, however, category feedback was given in the same way for both the target and distractor.

Importantly, in all conditions, categories designated as being part of the Distractor-Priority list produced classification accuracies that far exceeded chance performance, indicating that the inclusion of a category in a comparison, even if never as a target, is sufficient for significant learning gains to occur. This is consistent with prior research that suggests that simultaneous comparisons enhance the discriminability of presented items in a comparison (Gibson 1969; Mundy et al., 2007, 2009).

Our assessment results do not suggest a process by which participants became aware of which categories had appeared as targets more often and developed a response bias to guess those category labels more frequently. False alarms in the posttests did not differ by experimental condition. Conversely, increased hit rates for categories prioritized as targets, with the absence of increased false alarms, indicate true improvements in category perception.

What might explain the target-distractor effects we observed? One possibility is that feedback may cause preferential encoding or retention of relevant stimulus information that has just been used to make a classification. When a trial is correctly answered, a process similar to reinforcement learning may strengthen the tendencies to select and weigh more heavily information just used to make that classification. When an error occurs, feedback may initiate signals that downweight the information relied upon. Interestingly, our results suggest that these processes are centered upon the category queried on a given trial, despite the fact that feedback always provided correct labeling for both members of the comparison presented. Classic work in this area (Gibson, 1969) suggested that perceptual learning can be described as coming to selectively extract distinguishing features -- features and relations that make the difference between one category and another. One might think that learning of distinguishing features would be an inherently symmetrical process: Learning the stimulus properties that make exemplars of Category A different from those of Category B would seem to be the same as learning the properties that distinguish B from A. An intriguing possibility compatible with the present results is that the relation may not be symmetrical. Given the way a comparison task is posed, a learning experience framed as "Choose the image that comes from Category A" may preferentially benefit learning to distinguish A exemplars from others more so than the reverse.

An alternative explanation is that learners may prioritize attention to feedback given for the target category, given its framing as the goal of the trial. Consequently, although learners may come to pick up distinguishing features for the distractor category in each comparison, if they fail to regularly attend to the category label presented for distractors, they may struggle to integrate what they learn across separate trials. In other words, the comparison of target category A and distractor category B provide learning gains for both A and B, and a subsequent comparison between target category C and distractor category B provide similar gains, but if a learner does not come to recognize that both comparisons contained an instance of B, then they may fail to integrate the distinguishing features they have extracted in each comparison.

The results of this experiment advance our understanding of how paired comparisons support the learning of multiple perceptual classifications, as well as inspire other future directions. An interesting follow up question may be in regard to what would happen to the learning of distractor categories if the training phase had continued. Is learning simply slower for the distractor items, or is there a ceiling for how much can be learned about a distractor that would prevent it from ever reaching the same strength as categories presented as a target? Additionally, one could investigate how these results may interact with category difficulty; in cases where some categories may be more difficult to acquire than others–it may be advantageous to structure learning to prioritize the more difficult categories as targets.

Conclusion

Given the success of paired comparison trials to produce meaningful learning improvements in difficult real-world domains like skin lesion differentiation, it is important that we understand how learning is advanced for the presented categories. Both the target and distractor of a paired comparison trial benefit from the paired comparison learning format,

though learning gains can be asymmetric, providing an advantage for the target item. These results suggest that the task invoked on a paired comparison trial may be more specific than simply differentiating between two presented categories; rather, the task may be to find the information that allows for differentiation of the target category specifically.

Experiment 4

In Experiment 4, I shifted emphasis away from evaluating paired comparison's effects through classification accuracy, and instead aimed to more directly characterize the underlying perceptual changes that allowed for this learning to occur. As discussed in Chapter 1, prior work has established a clear relationship between our perception and our category representations, with perception dynamically updating to accommodate the demands of the task at hand (e.g., Goldstone, 2000, 2001; Hock et al., 1987; Schyns & Murphy, 1994). In particular, empirical results have shown evidence of changes in the perceived similarity structure of categories following categorization training (e.g., Livingston et al., 1998). Similarity updates such that items within the same category come to be perceived as more similar to one another, resulting in a phenomenon termed within-category compression or *acquired equivalence*. At the same time, items from different categories come to be perceived as more dissimilar, referred to as between-category expansion or *acquired distinctiveness* (Goldstone, 1994; Goldstone et al., 2001).

What drives these changes? As detailed by Gibson (1969), mechanisms of perceptual learning emphasize the discovery of stimulus features and attributes most important for differentiating one category from another. This heightened sensitivity to distinguishing information leads to the differences between categories becoming more pronounced, and objectively similar features may be viewed as increasingly dissimilar as a result. Further, the

commonalities that may exist between items of different categories may be down-weighted or ignored as this shared information does little to advance in the goal of differentiation. Within-category compression occurs as individuals learn to identify the commonalities among items within the same category while also suppressing non-critical variation. Together, these changes allow for quicker and more efficient interpretation and classification of visual stimuli.

Evidence of representational change has been measured through various methodological approaches. Psychophysical measures have been used to compare the ability to discriminate along category-relevant versus category-irrelevant boundaries following categorization training (Goldstone, 1994; Folstein et al., 2012, 2014), often reporting a heightened ability to discriminate at relevant category boundaries. Recently, this approach has also been paired with neuroimaging methods to provide additional insights into corresponding changes in brain activity (e.g., Folstein et al., 2013; 2015). Another common approach involves measuring the similarity of items, in which subjective reports of category similarity are solicited before and after categorization training to measure changes in perceived similarity structure (Ashby et al., 2023; Goldstone et al., 2001; Livingston et al., 1998; Reppa & Pothos, 2013).

Regardless of the general approach used, the categorization training in previous studies has predominantly employed single-item classification trials. Here, we looked to characterize the representational change that may occur following comparison-based training. Although categorization training by either approach both work toward the same end result (the discovery and preferential selection of the most critical perceptual patterns and features), the emphasis on category discrimination in the paired comparison approach raises a few interesting questions. If every trial primarily encourages the selection of distinguishing information, would representational change be driven primarily by between-category expansion? Similarly, how well

does one come to discover the commonalities that drive acquired equivalence among instances of the same category in this format?

A secondary focus of this experiment pertains to the characteristics of the categories being studied. In addition to using a specific type of categorization training, the majority of prior work has also concentrated on the learning and discrimination of a limited number of categories, typically only two. Further, the stimuli used are most often either artificial or naturalistic stimuli that have been artificially manipulated, such as the creation of categories through morphing different faces. These approaches offer clear advantages: the critical features and/or rules for categorization can be definitively established, and the objective similarity of items and categories can be directly measured and manipulated; however, they also may lack ecological validity as a result.

In real-world contexts, we frequently need to identify objects as belonging to multiple possible groups. Additionally, the features and dimensions most crucial for differentiation are likely to vary between different categories. For example, if we consider the skin lesion categories learned in the present work, one might use color information to help differentiate between the categories of Haemangioma and Solar Lentigo, but would find texture to be far more informative than color when differentiating between Solar Lentigo and Seborrheic Keratosis. Moreover, as previously discussed, most categories are likely to be considered ill-defined (e.g., Rosch, 1973), suggesting that categorization judgments based on a single feature or dimension will often be insufficient and rather depend upon the integration of information across multiple features and dimensions simultaneously. Given this, if tasked with learning multiple, ill-defined classifications, category representations will need to update in ways that accommodate more complex patterns of information.

We set two primary goals for the present study. First, we aimed to characterize the perceptual changes promoted by paired comparison categorization training. The second objective was to provide evidence for representational change in a paradigm that used a large number of complex, naturally occurring categories. To accomplish this, we collected similarity ratings from medically-naive participants for items belonging to ten skin lesion categories.

Participants viewed a pair of images containing either two images from the same category or one image from two different categories before being asked to rank how similar the depicted lesions were to each other. Similarity ratings were solicited as the first task of the experiment before participants were exposed to any category information, as well as collected again following categorization training with paired comparison learning trials. This approach enabled us to test for changes in similarity among exemplars from the same category, as well as for changes across different categories. Changes were evaluated globally by looking at the average similarity change for all same-category pairings and all different-category pairings, as well as at the individual category level (e.g., evaluating how similarity changed for all trials including Nodular Melanoma versus for trials containing Basal Cell Carcinoma).

Although commonly used in prior work, some research has suggested that similarity rating tasks may be influenced by task demands. Goldstone and colleagues (2001) proposed two possible sources for observed differences in similarity ratings. The first is the Changed Object Description Account, in which category learning alters the description of the objects themselves, representing true representational change. The second is the Strategic Judgment Bias Account, in which participants modify their similarity judgments to align with the expectation that items sharing a category label should be judged as more similar and items with different labels should be judged as more dissimilar.

Utilizing a paradigm where the difference between similarity ratings of learned faces relative to a neutral, uncategorized face were compared, Goldstone et al. (2001) provided evidence for within-category compression that could not be explained by the presence of a category label. Namely, when required to make a similarity judgment relative to an unlearned face, participants gave more similar answers for comparisons involving exemplars from the same category after training than before. However, evidence for between-category expansion could not be observed in the same way, the difference between ratings from items from different categories relative to the unlearned face was unchanged, suggesting that changes in similarity ratings between items of two learned categories could be the result of task demands.

Recently, another study attempted to quantify the proportion of changes in similarity ratings attributable to the changed object description account versus the strategic judgment bias account. Ashby and colleagues (2023) trained participants to learn categories of faces manipulated to follow either a similarity-consistent or similarity-inconsistent structure. In the similarity-consistent structure, faces within a category shared physical features, while in the similarity-inconsistent structure, category membership was orthogonal to physical features. Differences in similarity judgments before and after training indicated a category bias when the similarity structure was consistent; however, the presence of a common category label without a consistent underlying similarity structure failed to produce notable changes. The results of this study suggest that while a strategic judgment bias may have some influence on changes in similarity ratings, the effect is strongly driven by true perceptual change.

Here, we decided to test for possible effects of a strategic judgment bias account among categories that shared significant overlap in their associated label. In particular, within our category set are four category pairings that share overlap in their name (e.g., Basal Cell

Carcinoma and Squamous Cell Carcinoma). If learners are sensitive to the labels associated with the categories, as opposed to perceptual information contained within, then we might expect that categories that are similarly labeled would be judged to be more similar to each other after the category labels are learned in training, or, at the least, would not decrease in similarity to the same extent that non-similarly labeled categories (e.g., Basal Cell Carcinoma vs. Wart) might. In addition to looking at similarity change at the level of the specific pairing type and the individual category level, we also planned to evaluate similarity change with respect to label similarity.

Method

Participants

49 undergraduate psychology students from the University of California, Los Angeles were recruited. Participants provided informed consent and were awarded partial course credit for their participation.

Materials

Stimuli consisted of dermoscopic images of 10 skin lesion categories. The number of instances available varied per category with each category containing anywhere from 21 to 136 unique images, for a total of 690 available items. Two items per category were set aside to be used as novel items in the classification assessment.

Consistent with Experiments 2 and 3, the specific categories included: Actinic Keratosis, Basal Cell Carcinoma, Benign Nevus, Haemangioma, Lentigo Maligna Melanoma, Nodular Melanoma, Seborrheic Keratosis, Solar Lentigo, Squamous Cell Carcinoma, and Wart. Of note, included within these categories are four pairings that share significant overlap in their name. These are: Actinic Keratosis and Seborrheic Keratosis, Basal Cell Carcinoma and Squamous Cell Carcinoma, Lentigo Maligna Melanoma and Nodular Melanoma, and Lentigo Maligna

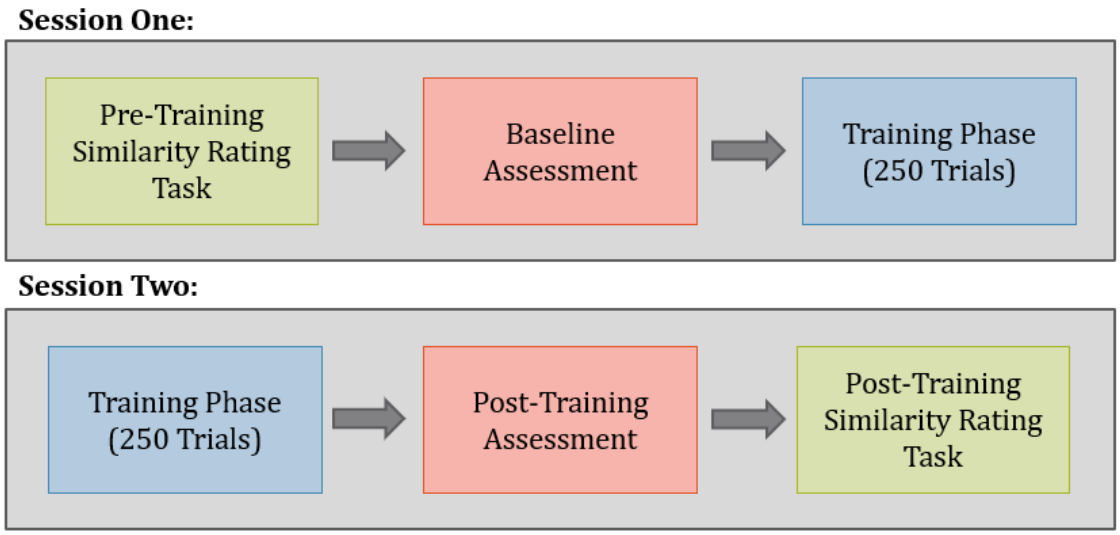
Melanoma and Solar Lentigo. These four pairings will be referred to as *similarly labeled categories*. All other different-category pairings will be designated as *non-similarly labeled categories*.

Design & Procedure

This study utilized a within-subjects design. An overview of the procedure can be seen in Figure 13. The experiment was divided into two sessions completed two days apart. In Session 1, participants completed a pre-training similarity rating task followed by a baseline assessment, and the first half of paired comparisons training. Session 2 contained the second half of training, a post-training assessment, and a post-training similarity rating task.

Figure 13

Procedure Overview (Experiment 4)



Note. Participants complete three different tasks (a similarity rating task, classification assessment, and training phase) twice across two experimental sessions. Experimental sessions were conducted 2 days apart.

Similarity Rating Task

The purpose of the similarity rating task was to measure perceived similarity between items from the same category as well as between items of different categories. At the start of each trial, a blank screen would be presented for 500ms before presenting two images side by side for 4000ms. The images would then be removed and participants would be asked to rank the similarity of the lesions that had been presented. A number line with options from 1 - 9 was then displayed. The labels “Most Dissimilar” and “Most Similar” were displayed on the extreme ends of the scale. Participants indicated their answer by clicking directly on a number. A higher number indicated a greater degree of similarity. An example trial is shown in Figure 14.

Figure 14
Similarity Rating Task Trial

Note. In a similarity rating trial, two images were presented side by side for 4000ms (left) before being removed and replaced by the prompt “Please rank the similarity of these lesions” and a 1-9 number line (right).

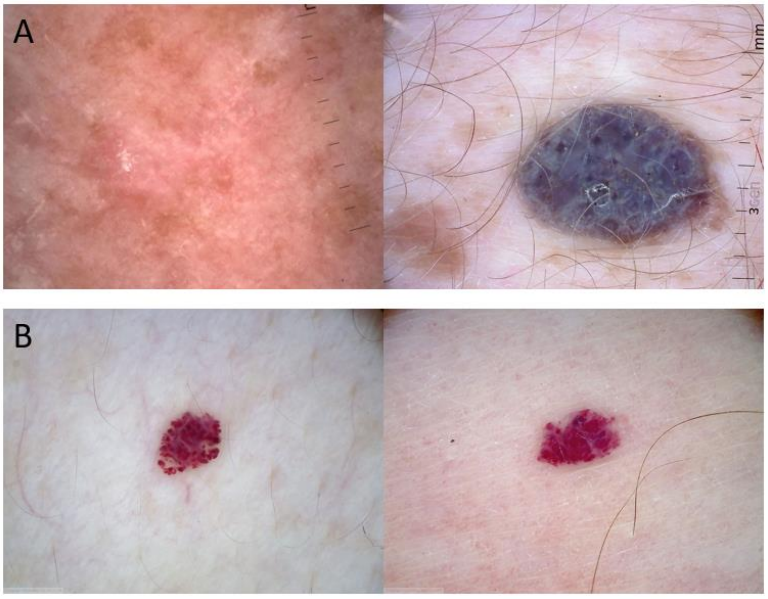
All participants completed six practice trials. These trials were intended to provide participants with a feel for the task and provide some exposure to the range of stimuli that they would encounter. The practice trials were selected to include two low-similarity pairings, two intermediate-similarity pairings, and two high-similarity pairings. An example of a low- and

high-similarity pairing can be seen in Figure 15. These image pairs were selected by the experimenter and independently rated by two research assistants familiar with the range of stimuli. Each category was shown at least once in practice. Participants were not given any type of feedback or category information.

Following the practice phase, participants completed 168 similarity rating trials. On these trials, participants were exposed to same-category and different-category pairings. Each category was paired with itself three times, for a total of 30 same-category pairings, and each category was also paired with an image from each other category three times[2], for a total of 135 different-category pairings. Finally, three attention check pairings, in which two identical images were displayed together, were also included.

Figure 15

Example Low- and High-Similarity Pairings



Note. Panel A depicts two low-similarity lesions paired together in the practice similarity rating task. Panel B depicts two high-similarity lesions.

[2] Due to a technical error, the Basal Cell Carcinoma and Solar Lentigo were paired together only twice and Seborrheic Keratosis and Solar Lentigo were paired together four times.

To avoid testing only exemplars where common and distinguishing features were easiest to discover and select, we included exemplars that encompassed a wide range of appearances for each category. A subset of learning trial data from the classification conditions in Experiment 2 was used to aid in selecting the specific images included here. Looking at the average accuracy for each presented learning item, we were able to determine item difficulty and infer category typicality for each image. We then selected images that spanned a wide range of difficulty within each category, with the exception that items that had been consistently misclassified at a high rate (average accuracy less than 20%) were disregarded for concern that they may be too ambiguous or disjointed for medically-naive learners to adequately learn.

Further, for each different-category pairing, the images selected were those previously shown to be mistaken for the accompanying category in the pair. When an item was misclassified in Experiment 2, the incorrectly chosen category label was recorded. Looking at these misclassifications, we chose items that demonstrated confusion with their paired category. For example, a pairing between Nodular Melanoma and Wart would contain one image of a Nodular Melanoma that had been repeatedly misclassified as a Wart and one image of a Wart that had been misclassified as a Nodular Melanoma. This approach allowed us to test for changes in perception in instances where it would be most needed.

Based on the availability of images for each category, and the criteria described above, the similarity rating task included 217 unique images. All trials were presented in a randomized order. Participants were given the opportunity to take a short break at the midway point. At no point during the task were participants given any feedback, informed of the ratio of same-category to different-category pairings, or provided with any category label information. The

similarity rating task was administered twice: once as the first task of the experiment and once as the last task of the experiment.

Training Phase

Following the pre-training similarity task, participants completed 500 paired comparison trials split equally across two training sessions. The paired comparison trials were identical in layout and function to those used in Experiments 1-3.

All categories were shown 100 times in the training phase, including 50 times as the target of a trial and 50 times as a distractor of a trial. The order in which categories were presented was randomized. Further, the pairing of categories on a trial (i.e., the category presented as the target and the category presented as the distractor) was also randomized; each category was able to be paired against an exemplar from any of the remaining nine categories on any trial.

Feedback was provided immediately after each trial and indicated whether the answer was correct/incorrect, as well as labeled both presented images. Participants were given 30 seconds to complete each learning trial and up to 10 seconds to view feedback. The training phase was the only portion of the experiment that provided any information about category membership.

Classification Assessments

Participants completed an assessment consisting of single-item classification trials at two time points: immediately before the training phase (baseline assessment) and immediately after the training phase (post-training assessment). The structure of the assessments was consistent with those used in Experiments 1-3. Each assessment contained two items per category that were

never shown in any other phase of the experiment, totaling 20 items. Participants had 30 seconds to complete each trial. No feedback was provided.

Exclusion Criteria

Only data from participants who completed both sessions of the experiment were included in the following analyses. Participants were excluded after data collection if they failed to demonstrate significant learning of the skin lesion classifications, measured as the difference in accuracy from the baseline assessment to the post-training assessment; in total, 9 participants were dropped for failing to improve their assessment score by at least 30%. Participants were also excluded if they answered with anything below an 8 (out of 9) on the attention check trials in the similarity rating task. An additional 9 participants were excluded for failing attention check trials, leaving a final total of 31 participants for analysis.

Dependent Measures and Data Analyses

For each participant, performance on the baseline assessment and post-training assessment, measured as the percentage of items correctly classified, was recorded. For the similarity rating task, ratings were averaged across same-category pairings and different-category pairings for each participant at both phases of the experiment. Contrasts were planned between the average rating of same- and different-category pairings at both phases of the experiment, as well as between pairings of the same type at each phase.

Assessment accuracy and similarity ratings for same-category pairings and different-category pairings were also calculated separately for each category. Correlation analyses were conducted to investigate possible relationships between perceived similarity and assessment performance. Finally, differences in reported similarity were compared between similarly labeled pairings and non-similarly labeled pairings.

Condition differences were evaluated using standard parametric measures with alpha set at .05. Effect sizes are reported for each difference.

Results

Training and Assessment Accuracy

In the training phase, participants completed 500 paired comparison learning trials with an average overall accuracy of 83.19% (*SD* = 4.96). The average assessment score before training was 17.26% (*SD* = 10.87) and grew to 72.74% (*SD* = 11.75) at the post-training assessment. Average accuracy for each of the ten categories is listed in Table 1.

Table 1
Assessment Accuracy by Category

	Assessment Accuracy	
Category	Pre-Training	Post-Training
Actinic Keratosis	0.16	0.82
Basal Cell Carcinoma	0.19	0.37
Benign Nevus	0.13	0.92
Haemangioma	0.21	1.00
Lentigo Maligna Melanoma	0.26	0.63
Nodular Melanoma	0.11	0.95
Seborrheic Keratosis	0.16	0.40
Solar Lentigo	0.21	0.65
Squamous Cell Carcinoma	0.15	0.77
Wart	0.15	0.76

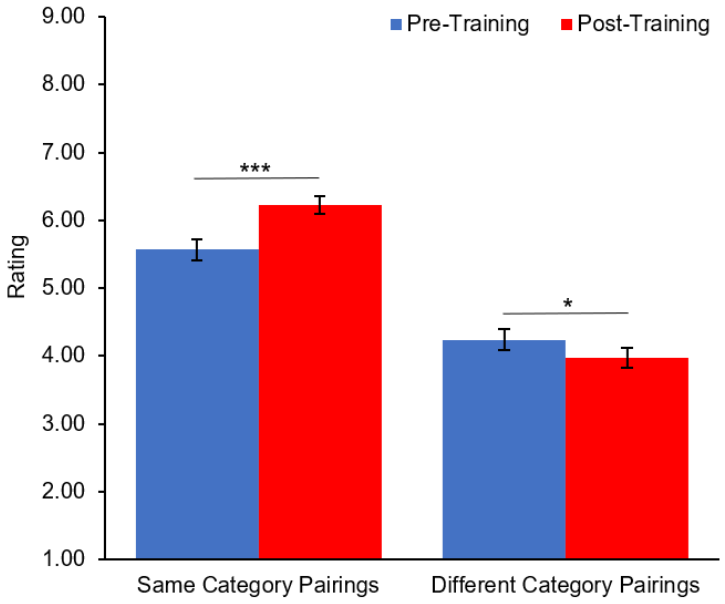
Note. Average accuracy, measured as proportion correct, is reported for each category at the pre-training and post-training assessments.

Similarity Ratings

Figure 16 displays the average similarity rating for same-category pairings and different-category pairings at the pre-training and post-training similarity rating tasks. A repeated measures ANOVA, with factors of pairing type (same-category vs. different-category) and phase (pre-training vs. post-training) was conducted on the similarity rating data. Analyses revealed a significant main effect of comparison type, $F(1, 30) = 376.04$, $p < .001$, $\eta_p^2 = 0.93$, a significant main effect of phase, $F(1, 30) = 5.13$, $p = .031$, $\eta_p^2 = 0.15$, and a significant comparison type by phase interaction, $F(1, 30) = 73.04$, $p < .001$, $\eta_p^2 = 0.71$.

Figure 16

Average Similarity Ratings



Note. Average similarity ratings given on the similarity rating task are graphed for same-category pairings and different-category pairings in the pre-training rating task and the post-training rating task. Statistical significance is indicated with asterisks: * $p < .05$, ** $p < .01$, ***$p < .001$.

Breaking down this interaction, at the pre-training similarity rating task, the average rating given for same-category pairings (M = 5.57, SD = 0.86) was significantly higher than the average rating given for different-category pairings (M = 4.24, SD = 0.72), $t(30) = 17.66$, $p < .001$, $d = 1.68$. The same pattern held at the post-training similarity task, with same-category pairings (M = 6.23, SD = 0.87) rated as significantly more similar than different-category pairings (M = 3.97, SD = 0.84), $t(30) = 17.14$, $p < .001$, with a greater effect size than observed at the pre-training task, $d = 2.64$.

For same-category pairings, the average similarity rating increased 0.67 points (SD = 0.60) from the pre to post-training rating task. This was confirmed to be statistically significant, $t(30) = 6.22$, $p < .001$, $d = 0.76$. For different-category pairings, the average similarity rating decreased by 0.26 points (SD = 0.56) from pre- to post-training. This was also found to be a statistically reliable difference, $t(30) = -2.62$, $p = .014$, $d = 0.35$.

Representational Change by Category

While the similarity rating results provide evidence for acquired equivalence for same-category pairings and acquired distinctiveness for different-category pairings overall, we were also interested in whether these perceptual changes impacted all categories in the set equally. We calculated a similarity score for same- and different-category pairings for each category. For same-category pairings, we took the average of the trials that contained images from only that category. For example, the same-category similarity score for the Wart category is the average of all trials that paired two Wart exemplars together. For different-category pairings, we took the average of every trial that contained one exemplar from that category. For Wart, this would include every trial that (only) one Wart exemplar included in, regardless of the other category displayed. Each category was involved in three same-category pairings and 27 different-category

pairings at each phase of the task. Table 2 contains the average similarity change score for each category, measured as post-training similarity minus pre-training similarity, divided between same-category pairings and different-category pairings.

Table 2

Changes in Similarity Rating by Category

Category	*Same Category Pairings*	*Different Category Pairings*
Actinic Keratosis	+0.76	-0.39
Basal Cell Carcinoma	-0.37	-0.28
Benign Nevus	0.00	-0.59
Haemangioma	+2.21	-0.20
Lentigo Maligna Melanoma	-0.03	-0.29
Nodular Melanoma	+1.73	0.00
Seborrheic Keratosis	-0.11	-0.06
Solar Lentigo	-0.27	-0.26
Squamous Cell Carcinoma	+1.21	-0.32
Wart	+1.55	-0.27

Note. Change in similarity score (post-training similarity - pre-training similarity) is reported for each category. Similarity change is averaged across all trials containing that category for each participant.

Different categories were affected differently by training. The observed similarity differences from the pre- to post-training for same-category pairings ranged from increasing by 2.21 points (Haemangioma) to decreasing by 0.37 points (Basal Cell Carcinoma). For different-category pairings, the changes ranged from decreasing similarity by 0.59 points (Benign Nevus) to remaining unchanged (0 point difference; Nodular Melanoma). To find an explanation for these differences, correlation analyses were run between pre-learning similarity ratings, post-

learning similarity ratings, similarity difference scores, and posttest classification accuracy for each category.

First, to check whether the differences in similarity change among categories could be attributed to ceiling or floor effects, such that the same-category pairings that were most similar before training did not have room to become more similar after training or the different-category pairings that were the least similar before training did not have room to become less similar after training, we ran a correlation analysis between pre-training similarity and the similarity difference score separately for the same-category pairings and the different-category pairings. No relationship was found between pre-training rating and similarity change for same-category pairings, $r(8) = -.18$, $p = .630$, or for different-category pairings, $r(8) = -.52$, $p = .121$.

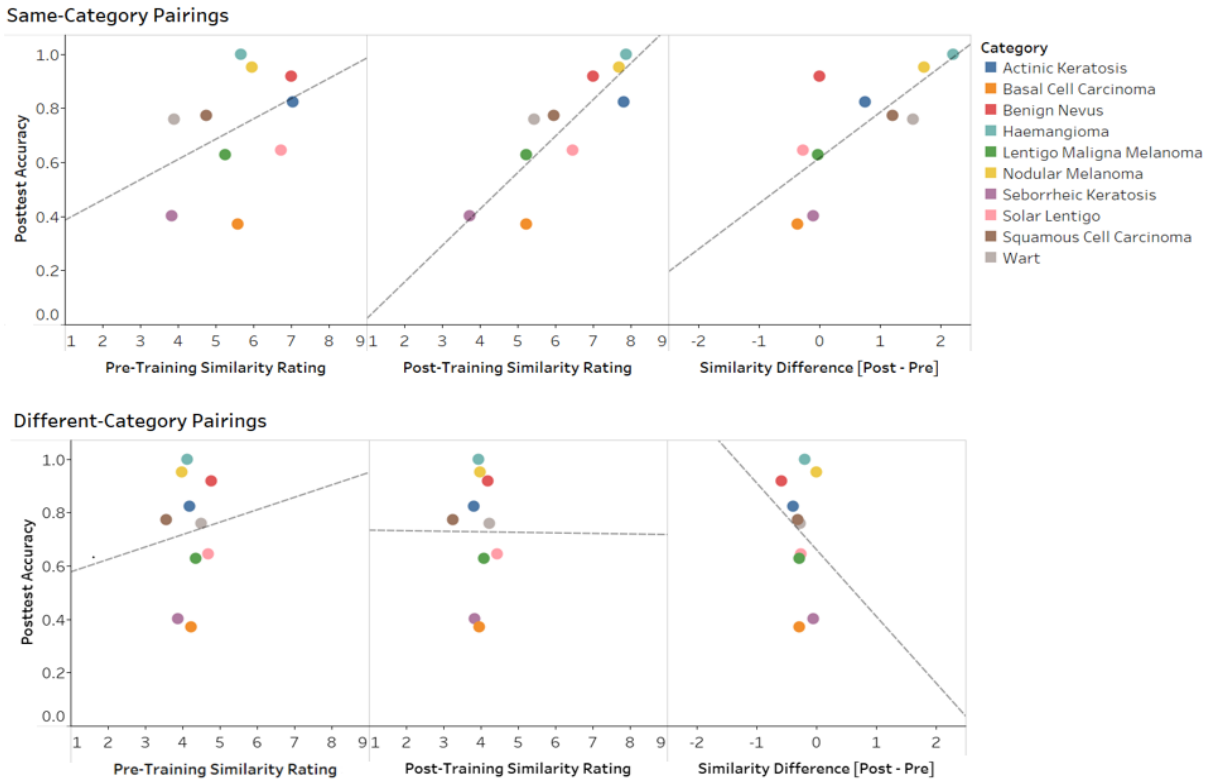
Next, we investigated whether perceived similarity correlated with the learner's performance on the classification assessment following training. Scatterplots of this data can be seen in Figure 17. Initial pre-training similarity ratings were not found to have a reliable relationship with posttest accuracy for either same-category pairings, $r(8) = .41$, $p = .246$, or for different-category pairings, $r(8) = .08$, $p = .826$. Looking at post-training ratings, we find a strong positive correlation between the average similarity rating and posttest accuracy, $r(8) = .85$, $p = .002$, such that the more similar items from the same-category were rated to be the more accurate participants were in classifying novel instances from that category. There was no relationship found for different-category pairings, $r(8) = .00$, $p = .993$.

Finally, we also tested for a relationship between similarity difference scores and posttest accuracy. Mirroring the results of the post-training similarity correlations, a significant positive relationship was found between same-category difference scores and posttest accuracy, such that the categories that increased most in perceived similarity following training were also more often

correctly classified on the post-training assessment, $r(8) = .74$, $p = .015$. No relationship was found for different-category pairings, $r(8) = -.19$, $p = .602$.

Figure 17

Similarity Ratings by Assessment Accuracy



Note. The relationship between average posttest accuracy (measured as proportion correct) and average pre-training similarity rating, post-training similarity rating, and similarity rating difference are depicted for same-category pairings (top) and different-category pairings (bottom).

Label Similarity

Table 3 lists the similarity change for each of the four similarly labeled category pairings relative to all other non-similarly labeled different-category pairings. The average similarity change for the similarly labeled categories ($M = -0.25$, $SD = 0.79$) was very similar to the average similarity change across all other different-category pairings ($M = -0.27$, $SD = 0.56$), and

a dependent samples t-test determined this difference to be unreliable, $t(30) = 0.17$, $p = .868$, $d = 0.03$.

Table 3

Similarity Rating Change for Similar and Non-Similar Category Pairings

Category Pairing	*Pre-Training Similarity*	*Post-Training Similarity*	*Similarity Change*
Actinic Keratosis vs. Seborrheic Keratosis	4.35	4.57	+0.22
Basal Cell Carcinoma vs. Squamous Cell Carcinoma	3.76	3.13	-0.63
Lentigo Maligna Melanoma vs. Nodular Melanoma	5.58	5.32	-0.26
Lentigo Maligna Melanoma vs. Solar Lentigo	6.69	6.35	-0.34
All Other Different-Category Pairings	4.15	3.88	-0.27

Note. Similarity ratings (pre and post-training) and similarity change scores are given for the four similarly labeled category pairings, as well as for the average of all other different-category pairings.

Discussion

The present study investigated whether and how representational change could occur as the result of comparison-based training in a relatively large, complex category domain. As expected, items from the same category were rated as more similar to each other than items from different categories, even before categorization training. Importantly, this difference was exaggerated following learning and our findings provide evidence of both within-category acquired equivalence and between-category acquired distinctiveness following comparison-based categorization training.

Notably, the effect of within-category compression was found to be significant with a medium effect size. This suggests that participants became more adept at recognizing the similarities among items within the same category, reducing the perceived differences between them. This finding is particularly important given the paired comparison trial format used in our study, which required category acquisition to occur solely through trials emphasizing the discrimination of items from different categories. However, shifting focus to an individual category level, we see that the magnitude of this effect was highly variable, with increases in similarity being observed for only five of 10 learned categories.

One possible explanation for this result has to do with the amount of variability that exists within each category. Within these categories, there exist natural differences in how similar items from the same category appear to each other, as well as how heavily they overlap with instances of other categories. Given the structure of paired comparison trials, attention is primarily devoted to finding the distinguishing information in the items presented, possibly making the discovery of commonalities and patterns across exemplars of the same category a less prioritized goal. Categories that are then characterized by high-variability or significant feature overlap with other categories in the set may be at a disadvantage for discovering uniting features and achieving acquired equivalence. The results from our correlation analyses showed a strong positive relationship between similarity change from pre- to post-training and assessment performance, as well as between post-training similarity and assessment performance, suggesting that the categories that did not show a compression effect were the categories that were learned the least. Given that perceptual learning enables changes in perception, which then enable category learning, it stands to reason that if the changes in perception did not adequately occur, then neither would the learning of the classification.

When looking at the change in similarity for different-category pairings, the overall effect of between-category expansion was smaller than that of within-category compression, though was much more consistent across categories with nine of the 10 showing an average decrease in similarity from pre to post and no categories yielding an overall increase in similarity. Interestingly, despite the emphasis on differentiation promoted by perceptual learning, changes in between-category similarity were not found to have any relationship with posttest performance. To explain why this might be the case, we look more closely at how the between-category expansion was quantified for each category.

We chose to average the different-category pairing similarity across all trials containing a given category, giving us a general metric of how similar a given category was perceived relative to all other categories in the set simultaneously; however, this approach may obfuscate smaller patterns and differences between specific different-category pairings. An alternative to this would be instead to observe differences at the level of the specific pairings (e.g., Nodular Melanoma-Wart) rather than at the category level. Analyzing the difference in similarity across all 45 different-category pairings, rather than across the 10 categories, reveals greater variation in similarity changes. In particular the most extreme pairs show a similarity difference ranging from -1.13 (Benign Nevus vs. Squamous Cell Carcinoma) to +0.98 (Nodular Melanoma vs. Squamous Cell Carcinoma). It may be the case that becoming attuned to the differences between certain category pairings, likely the most similar categories, is more important than coming to see each category as generally more dissimilar from all categories. Future work may benefit from an additional form of assessment that measures not just classification accuracy, but also one's ability to discriminate between specific pairings following training to further test this idea.

Finally, while we argue that the changes observed in the similarity scores are the direct result of changes in perception, it has been suggested that the use of category labels could also drive these differences. In particular, learners may adopt a task-demand, strategic bias account, in which a participant modifies their similarity judgments to align with expectation that items sharing a category label should be judged as more similar than categories with differing labels (Goldstone et al., 2001).

There are a few reasons to believe that, in the present work, similarity judgments were made based largely on the perceptual information contained in the images rather than in response to the category label. First, with participants exposed to over 500 different images throughout the experiment, the possibility that a learner memorized each image and its associated label before rating each pair in the similarity rating task is highly unlikely. This suggests that if label retrieval does occur before similarity ratings are given in the present work, it is reflective of the observer's selection of the uniting and/or distinguishing perceptual information contained within each image, despite not being a requirement of the task. This may be considered in contrast to the categories used in Goldstone et al., in which each of the two learned categories were only defined by two images each.

Furthermore, the post-training similarity scores and the magnitude of compression or expansion observed differed considerably by category and by pairing. If participants were simply reacting to whether labels matched or did not match, we would not expect such drastic differences among these measures. In particular, results from our correlational analyses demonstrated a clear relationship between similarity change and one’s ability to classify previously unseen, novel images from that category. This suggests a clear link between ability to

pick up on critical perceptual information in new stimuli and the perceived similarity of items in that category.

Finally, the results from our analyses on similarly labeled categories provide additional insight into the role of labels in learning. Four of the learned categories in the current set shared labels that indicated a relationship at another level, e.g., Nodular Melanoma and Lentigo Maligna Melanoma both belong to the more superordinate category, and shared name, of *melanoma*. If participants had adopted a strategic bias account and were relying upon some type of strategic cognitive processing to make their similarity judgements, then one might expect that the differences in similarity ratings among different-category pairings might be driven by differences in label overlap rather than featural overlap. Our results indicate that this was not the case and that category pairings with overlap in their associated labels showed a very similar decrease in similarity as categories with more dissimilar labels.

This finding is consistent with the work by Ashby and colleagues (2023), demonstrating that the presence of a uniting label in the absence of underlying similarity structure was insufficient to produce any meaningful changes in similarity reportings. While these results do not entirely isolate perceptual categorization from category labels, they provide further support for changes in similarity judgements driven by corresponding changes in the perceived similarity structure of category representations.

Conclusion

Categorization training using paired comparison learning trials can lead to significant changes in category representation, measured through subjective reports of similarity. Evidence of both acquired equivalence and acquired distinctiveness was found. This finding is crucial as it highlights the capacity of this contrast-focused learning approach to not only enhance the

discrimination between different categories but also aid in the discovery of commonalities within members of the same category. Importantly, our study demonstrates that representational change occurs even when dealing with a large number of ill-defined, naturally occurring categories. This work has important implications in advancing our understanding of how contrastive comparisons advance learning, as well as in understanding how perceptual learning underlies real-world classification tasks more generally.

CHAPTER 5

Adaptive Approaches to Comparison

Chapters 3 & 4 demonstrated that the effectiveness of a simultaneous comparison may depend on a number of factors including the task of the trial and the framing of each category; however, one factor that has not yet been considered is that the value of a given comparison may also vary from learner to learner. While the previous chapters ask questions about how a comparison may be structured to be most effective on a more global level, the introduction of adaptive learning methods allows us to focus on the individual in the present chapter.

As reviewed in Chapter 1, adaptive learning refers to methods and technologies that seek to personalize and optimize the learning process for an individual by adjusting events in learning based on the individual's responses or performance. In earlier work, we created and tested a new type of adaptive learning element designed to capitalize on the theorized benefits of learning with comparisons, which we call *adaptively triggered comparisons (ATCs)*. ATCs are comparison trials whose timing and content are prompted by errors in which a learner confuses two categories. Unlike ARTS which uses only accuracy and response times to structure learning events, ATCs consider not only whether an item was classified correctly, but also the specific answer that was given for each incorrectly classified trial. New learning events (comparison trials) are then triggered by identifying patterns in the incorrect answers given for each category.

To see if this type of adaptive element provides learning advantages beyond those already given by an effective adaptive system, we studied ATCs in the context of an adaptive learning session using a face-identification paradigm. If, during single-item classification trials, a learner was presented with a picture of David but responded "Alexander," that information could be used to generate a simultaneous, between-category comparison trial in which pictures of

David and Alexander appeared side by side. Intuitively, the value of ATCs might be twofold: tracking learners' errors identifies specific confusions during learning, while simultaneous comparisons may help resolve them by facilitating the discovery of critical invariants that may be harder to detect when items appear in isolation.

Across two experiments, we tested whether learning could be improved through the inclusion of ATCs. In both experiments, participants were trained to classify 22 different face categories. Experiment 1 compared performance using only single-item classification trials scheduled through ARTS (*No Comparisons Condition*) with a condition mixing single-item classification trials and ATC trials (*ATC Condition*). Experiment 2 asked whether the inclusion of adaptive comparison trials enhanced learning to a greater degree than learning with an equal number of non-adaptive simultaneous comparison trials (*Non-Adaptive Comparisons (NAC) Condition)*.

Preliminary Work: Adaptively Triggered Comparisons in Face Identification

Experiment 1: Method

Participants

82 undergraduate psychology students were recruited from the University of California, Los Angeles to participate. Two participants were dropped for failure to follow instructions and an additional four were dropped for technical issues during the experiment. The remaining participants were equally distributed between the ATC condition (n = 38) and the No Comparisons condition (n = 38). Participants provided informed consent and received partial course credit for their participation.

Stimuli

We used 22 human male faces with five distinct pictures of each person for a total of 110

unique images taken from a larger database. Four images of each of the 22 categories were used in the learning phase. The fifth image in each category was set aside for use as a novel stimulus in the posttests. Non–face details such as hairstyle or visible clothing varied across images within the same category. Background and final image size were the same for all images.

Each face category was identified with a name. The names were chosen to be unremarkable, and were taken from the Social Security list of most common names given in the United States in 2000-2009. The names and images used were identical to those used in Experiment 1 of this dissertation.

Design & Procedure

Participants were randomly assigned to either the ATC condition or No Comparisons condition. All participants completed a learning phase followed by an immediate posttest. Delayed posttests were completed one week later.

Figure 18 shows an example of the types of learning trials a participant could encounter. On most learning trials in either condition, a single image was presented with all 22 possible names shown alphabetically below. Participants were instructed to select the name that matched the presented picture. Participants had 20s to answer and were given immediate feedback. Feedback indicated whether their given answer was correct or incorrect and displayed the phrase "This is [Category Name]" at the bottom of the screen.

In the ATC condition, participants' answers on single-item learning trials were monitored for patterns that indicated systematic confusions. A confusion was defined as two incorrect answers that involved the same two categories. For example, for categories Alexander and Benjamin, a confusion would be registered if a) an Alexander exemplar was incorrectly labeled "Benjamin" twice; b) a Benjamin exemplar was incorrectly labeled "Alexander" twice; or c) an

Alexander exemplar was incorrectly labeled “Benjamin” once and a Benjamin exemplar was incorrectly labeled “Alexander" once. When a confusion was detected, it triggered the creation of a comparison trial in the ATC condition.

Comparison trials were set to appear at the next scheduled presentation of whichever confused category appeared soonest in learning. Given the way categories were sequenced, this resulted in ATC trials occurring after a delay of 3 trials from the second misclassification, with some “jitter,” such that the delay was sometimes 2 or 4 (25% each).

On a comparison trial, one randomly chosen exemplar from each of the two confused categories was presented with the prompt “Which is [Category Name]?” Participants were required to select the image that they believed matched the name provided. Feedback was given immediately with the appropriate category label displayed beneath each picture. Participants then resumed single-classification trials. Another comparison trial between the two categories would be triggered only if the categories were again confused on a subsequent single-classification trial.

Figure 18
Example Learning Trials: ATC Study

Note. The left panel depicts an example single-item classification learning trial (“Choose the name that goes with the face”). The right panel depicts an example comparison trial (“Which is [Category Name]?”).

In both conditions, categories were adaptively scheduled and interleaved through the Adaptive Response Time-Based Sequencing (ARTS) system. During learning, each category is assigned a priority score indicating the relative benefit of that category appearing on the next learning trial. Priority scores for a category are a function of learner accuracy, response times, trials elapsed since last presentation, and progress toward meeting mastery criteria (See Mettler et al., 2016 for computational details.) The sequencing algorithm presents the highest priority item on each trial. An enforced delay is also implemented: a category could not recur while feedback from a recent instance could still reside in working memory. In the present study, we used an enforced delay of 3 with a +/- 1 jitter. As an individual's learning strength for a given category increases (indicated by accuracy and lower RTs), the ARTS algorithm automatically generates lower priority, and longer recurrence intervals, as an inverse function of the logarithm of reaction time.

Participants continued learning trials until all categories reached mastery. Mastery criteria required four consecutive correct classifications, each given in under five seconds. By using mastery criteria, we can ensure that participants in both conditions are given as much practice with each category as needed to learn to the same strength. All categories remained in learning until the last category was mastered. Immediately following learning, participants completed a 44-item posttest including one previously seen and one novel exemplar per category, randomized and presented sequentially. Posttest trials were the same as single-classification learning trials; however, no feedback was given. A delayed posttest, administered one week later, was identical in content and structure to that of the immediate posttest.

Dependent Measures and Data Analyses

For each participant, the number of learning trials invested to achieve mastery for all categories was recorded. For the ATC condition, the total number of learning trials included both single-classification trials and comparison trials. Each comparison trial completed was counted as one trial invested; although these trials contained two images, participants supplied only one answer and the outcome of the trial could only advance progress toward the set mastery criteria for the category whose label was presented as the target of the trial. Time invested in learning was also recorded for each participant.

For each participant in the ATC condition, the number of times each particular category combination (e.g., Alexander-Benjamin) was triggered for a comparison trial was recorded and ranked by frequency. Kendall's coefficient of concordance (*W*) was used to measure the extent to which participants agreed on the most confusable category pairs. The *W* coefficient was then linearly transformed to reveal the average Spearman's correlation between all possible pairs of raters (Kendall & Gibbons, 1990).

At the posttests, classification accuracies were recorded for each participant. Posttest performance was compared at each timepoint (immediate posttest and delayed posttest) as well as divided between performance on items seen in learning (old) and unseen (novel) items. Our primary dependent variable was *learning efficiency* -- essentially a *rate* obtained by dividing accuracy gains by learning investments. Efficiency measures are useful when mastery criteria are used, as they incorporate into a single measure both variations in trials to criterion and posttest performances (e.g., Mettler & Kellman, 2014). In the present work, we calculated two types of efficiency scores based off of two different learning investments: trial-based efficiency and time-based efficiency. We multiplied our time-based efficiency scores by 100 to obtain a learning rate

with the resulting number giving the percent gain in accuracy at either the immediate or delayed posttest per 100 trials invested in learning. Time-based efficiencies were multiplied by 10 to reflect the percent gain in accuracy for every ten minutes invested in learning. Condition differences were evaluated using standard parametric measures with alpha set at .05.

Results

Learning Investments

The ATC condition required significantly fewer learning trials to reach mastery (M = 345.95, SD = 76.41) than did the No Comparisons condition (M = 405.58, SD = 126.73). This difference of about 15% in trials to criterion was reliable, $t(74) = 2.48$, $p = .015$, *Cohen's d* = 0.57. Time taken to reach the mastery criteria also significantly differed between condition with those in the ATC condition taking less time to reach mastery (M = 29.28 min., SD = 7.82) relative to the No Comparisons condition (M = 35.60 min., SD = 12.62), $t(74) = 2.63$, $p = .010$, d = 0.60**.**

Efficiency Measures

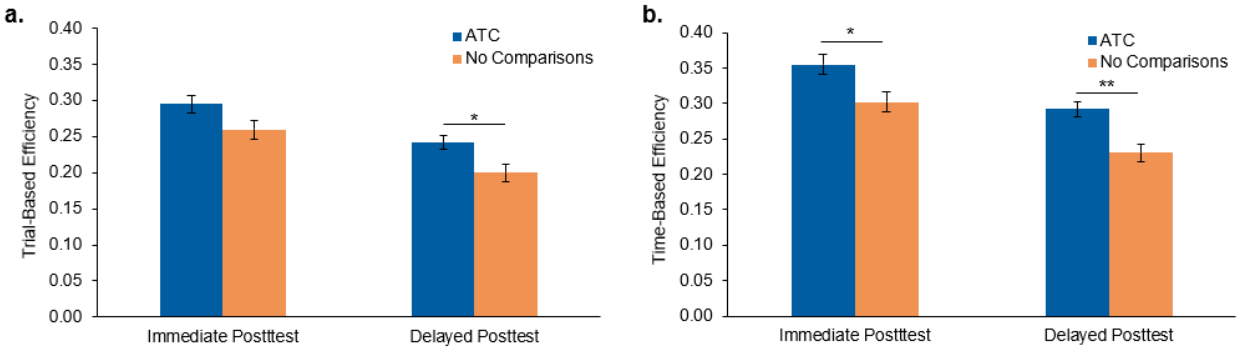
Trial-Based Efficiency

Figure 19a shows the mean trial-based efficiency scores for both conditions at each posttest phase. At the immediate posttest, the ATC condition yielded a higher efficiency (M =.0030, SD = .0007) than the No Comparisons condition (M = .0026, SD = .0008), such that for every 100 learning trials invested, the average gain in immediate posttest accuracy was 30% for participants in the ATC condition and 26% in the No Comparisons condition. This difference bordered on significance, $t(74) = 1.99$, $p = .050$, $d = 0.53$.

At the delayed posttest, efficiency reliably favored the ATC condition (M = .0024, SD = .0006) relative to the No Comparisons condition (M = .0020, SD = .0007), such that for every

100 learning trials invested, the average gain in delayed posttest accuracy was 24% in the ATC condition and 20% in the No Comparisons condition, $t(74) = 2.57$, $p = .012$, $d = 0.61$. No interaction between condition and posttest phase was found, $F(1, 74) = 0.53$, $p = .469$.

Figure 19
Efficiency Results (ATC Faces: Exp. 1)



Note. Immediate and delayed posttest efficiency scores are shown for the Adaptively Triggered Comparisons (ATC) condition and No Comparisons condition. Trial-based efficiency scores (a) indicate mean accuracy gain per 100 learning trials invested. Time-based efficiency scores (b) indicate mean accuracy gain per 10 minutes invested. Error bars represent +/- 1 standard error of the mean. Statistical significance is indicated with asterisks: * $p < .05$, ** $p < .01$, ***$p < .001$.

Time-Based Efficiency

Figure 19b shows the mean time-based efficiencies for both conditions at each posttest phase. As with trial-based efficiency, the ATC condition yielded a higher time-based efficiency (M =.036, SD = .009) than the No Comparisons condition (M = .030, SD = .009) at the immediate posttest, such that the accuracy gain for every 10 minutes invested in learning was 36% for those in the ATC condition and 30% for those in the No Comparisons condition, $t(74) = 2.58$, $p = .012$, $d = 0.67$.

At the delayed posttest, efficiency once again favored the ATC condition (M = .029, SD = .009) relative to the No Comparisons condition (M = .023, SD = .008), such that for every ten minutes invested the average gain in delayed posttest accuracy was 29% in the ATC condition

and 30% for participants in the No Comparisons condition, $t(74) = 3.27$, $p = .002$, $d = 0.70$. There was no reliable interaction between condition and posttest phase, $F(1, 74) = 0.50$, $p = 482$.

Accuracy Measures

At the posttests, all participants demonstrated an ability to classify the items previously seen in learning, as well as novel instances. Looking at assessment accuracy across all items, immediate posttest accuracy across all items did not differ between conditions (ATC: $M = 0.97$, SD = .03; No Comparisons: $M = 0.97$, $SD = .04$), t(74) = 0.16, $p = .875$, $d = 0$. Given that all participants learned to the same mastery criterion, we did not expect to see differences between conditions. Assessment accuracy was then divided into old and novel items and performance between conditions was compared. While old items were classified more accurately than novel items in both groups, there was no reliable evidence of a condition difference for classification accuracy on the old items (ATC: $M = 0.99$ $SD = .03$, No Comparisons: $M = 0.99$, $SD = .02$), $t(74) = 0.21$, $p = .839$, $d = 0.04$, or on novel items (ATC: $M = 0.96$, $SD = .05$, No Comparisons: $M = 0.96$, $SD = .06$), $t(74) = 0.10$, $p = .922$, $d = 0.02$.

At the delayed posttest, assessment accuracy was higher in the ATC condition ($M = 0.80$, $SD = .11$) than the No Comparisons condition ($M = 0.75$, $SD = .14$); however, this difference did not reach statistical significance, $t(74) = 1.61$, $p = .111$, $d = 0.39$. When the assessment was divided into old and novel items, classification accuracy favored the ATC condition for both item types, but neither difference reached significance (Old Items, ATC: $M = 0.81$, $SD = .11$, No Comparisons: $M = 0.75$, $SD = .16$, $t(74) = 1.59$, $p = .117$, $d = 0.37$; Novel Items, ATC: $M = 0.79$, $SD = .12$, No Comparisons: $M = 0.75$, $SD = .13$, $t(74) = 1.42$, $d = 0.33$).

Adaptive Comparisons

The average number of comparison trials triggered by a participant in the ATC Condition was 31.71 (*SD* = 15.78) including an average of 18.05 (*SD* = 6.14) different category combinations. Following their initial confusion, participants recurrently triggered the same two categories for comparison an average of 0.70 (*SD* = 0.38) more times throughout the course of learning.

A chi-square test of independence determined that there was reliable non-zero agreement in confusability rankings among participants, X^2 (230, N = 38) = 1293.93, $p < .001$. However, the coefficient of concordance indicated that the extent of this agreement was small, $W = 0.15$, and the resulting Spearman correlation coefficient revealed the relationship between participants' rankings to be weak, $r_s = 0.13$.

ATCs in Face Learning: Experiment 2

Results from Experiment 1 showed a benefit of learning with the inclusion of ATC trials. Provided with an opportunity to confront their confusions shortly after making them, participants were able to progress through the learning phase with fewer errors, reducing the amount of time and trials needed to achieve mastery, and without hindering performance in the short or long term.

From these results alone, it is not possible to determine whether the ATC benefit was due to the adaptive nature of the trials or simply to giving participants some opportunity to engage in simultaneous comparisons. To test whether the effect was at least in part due to the adaptive nature of the trials, Experiment 2 introduced a new *Non-Adaptive Comparisons (NAC)* condition in which participants engaged in roughly equal numbers of simultaneous comparison trials in learning as the ATC condition, but without regard to any specific confusions.

Method

Participants

62 undergraduate psychology students completed the experiment for course credit, however, two participants were excluded for failing to follow instructions. The remaining participants were equally split between the ATC and NAC conditions. Participants were awarded partial course-credit for their participation.

Design & Procedure

Participants were randomly assigned to either the ATC condition or NAC condition. The ATC condition was identical to that used in Experiment 1. For the NAC condition, we aimed to keep the proportion of comparison trials to single-target learning trials consistent with the proportion that naturally arises in the ATC condition. Based on Experiment 1 and additional pilot data, we estimated that an ATC participant performing at the average in terms of learning trials invested would encounter approximately one comparison trial for every 8 single-item classification trials. Using this, the NAC condition was set to insert a comparison trial on every ninth learning trial.

The categories in the NAC condition were not based on learning confusions. Instead, whichever category was due to show up after the eighth single-classification trial (as determined by the ARTS sequencing) was then included in the comparison trial as the target, and the competing exemplar was chosen randomly from all other 21 categories.

All participants continued in the learning phase until all categories were mastered. However, rather than retaining all categories in the learning phase from start to finish, the learning phase was adjusted such that once a category was mastered, it was retired from the learning set. Retired categories only re-emerged when necessary as filler items to achieve correct

spacing intervals for the remaining, unmastered categories. This decision was made to address concerns of overlearning for categories mastered early on, as large numbers of participants achieved ceiling performance on the immediate posttest (n = 33) and delayed posttest (n = 3) in Experiment 1. Following learning, participants completed an immediate and one-week delayed posttest.

Results

Learning Investments

The ATC condition required significantly fewer learning trials to reach the mastery criteria (M = 256.30, SD = 65.38) than the NAC condition (M = 329.17, SD = 90.81), $t(58)$ = 3.57; p = .001, d = 0.92. Time taken to reach the mastery criteria was also reliably lesser for participants in the ATC condition (M = 26.67 min., SD = 9.54) than in the NAC condition (M = 34.27 min., SD = 11.81), $t(58)$ = 2.74, p = .008, d = 0.71.

Efficiency Measures

Trial-Based Efficiency

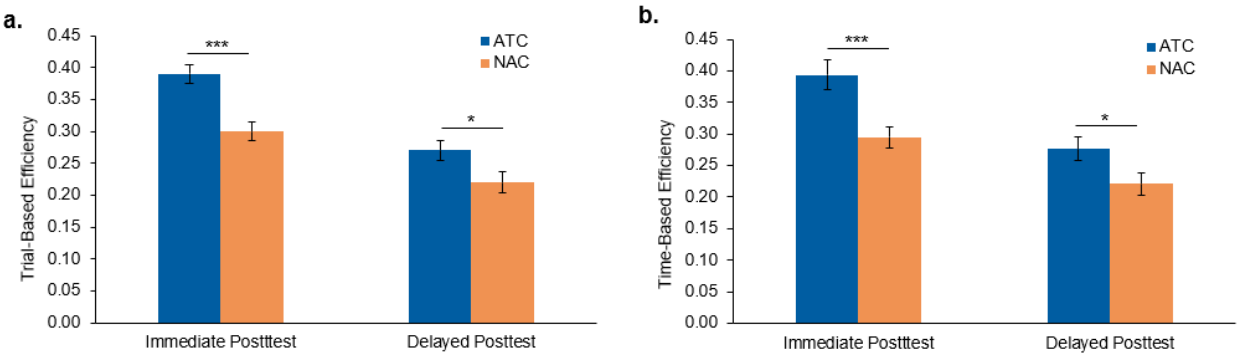
Efficiency measures were calculated for each participant for both posttest phases, and mean efficiencies were compared. Figure 20a shows the average trial-based efficiency by condition and posttest phase. The ATC condition yielded higher efficiency at both posttests. At the immediate posttest, participants in the ATC condition saw an average accuracy increase of 39% per 100 learning trials invested (M =.0039, SD = .0008). Those in the NAC condition saw a 30% increase (M = .0030, SD = .0008). An independent t-test determined this difference to be highly reliable, with a large effect size, $t(58)$ = 4.21; p < .001, d = 1.13.

At the delayed posttest, efficiency once again reliably favored the ATC condition with an average accuracy increase of 27% per 100 learning trials invested (M = .0027, SD = .0009)

relative to the NAC condition which yielded an accuracy gain of 22% per 100 trials invested, (*M* = .0022, *SD* = .0009), *t*(58) = 2.19; *p* = .033, *d* = 0.56.

Figure 20

Efficiency Results (ATC Faces: Exp. 2)



Note. Immediate and delayed posttest efficiency scores are shown for the Adaptively Triggered Comparisons (ATC) condition and Non-Adaptive Comparisons (NAC) condition. Trial-based efficiency scores (a) indicate mean accuracy gain per 100 learning trials invested. Time-based efficiency scores (b) indicate mean accuracy gain per 10 minutes invested. Error bars represent +/- 1 standard error of the mean. Statistical significance is indicated with asterisks: * $p < .05$, ** $p < .01$, ***$p < .001$.

Time-Based Efficiency

Figure 20b depicts the average time-based efficiency across conditions and posttest phases. At the immediate posttest, time-based efficiency was greater for the ATC condition (*M* = 039, *SD* = .013) than the NAC condition (*M* = .029, *SD* = .009), such that for every ten minutes invested in learning, immediate posttest accuracy improved 39% for the ATC condition and only 29% for the NAC condition. An independent samples t-test determined this difference to be reliable with a large effect size, *t*(58) = 3.49, *p* < .001, *d* = 0.91.

At the delayed posttest, time-based efficiency also reliably favored the ATC condition over the NAC condition, such that for every ten minutes invested in learning, accuracy on the

delayed posttest increased 28% for those in the ATC condition and only 22% for those in the NAC condition, $t(58) = 2.09$, $p = .041$, $d = 0.55$.

Accuracy Measures

All participants demonstrated an ability to classify both old and novel exemplars on the posttest assessments. Immediate posttest accuracy across all items reliably varied by condition with those in the ATC condition scoring higher ($M = 0.94$, $SD = .05$) than those in the NAC condition ($M = 0.91$, $SD = .06$), $t(58) = 2.03$; $p = .047$, $d = 0.54$. When divided into old and novel items, no reliable difference was found between condition for performance classifying old items (ATC: $M = 0.96$, $SD = .06$, NAC: $M = 0.94$, $SD = .05$), $t(58) = 1.07$, $p = .287$, $d = 0.27$; however, accuracy on novel items was found to significantly differ between conditions, such that those in the ATC condition classified novel items more accurately ($M = 0.93$, $SD = .06$) than those in the NAC condition ($M = 0.88$, $SD = .08$), $t(58) = 2.41$, $p = .019$, $d = 0.63$. At the delayed posttest, accuracy across all items did not differ significantly between conditions (ATC: $M = 0.66$, $SD = .14$; NAC: $M = .67$, $SD = .18$), $t(58) = 0.13$, $p = .898$, $d = 0.06$. There was also no significant difference between conditions when the assessment was divided into old items only (ATC: $M = 0.70$, $SD = .15$, NAC: $M = 0.68$, $SD = .18$), $t(58) = 0.25$, $p = .805$, $d = 0.07$, or novel items only (ATC: $M = 0.63$, $SD = .15$, NAC: $M = 0.65$, $SD = .19$), $t(58) = 0.48$, $p = .633$, $d = 0.12$.

Adaptive Comparisons

On average, participants in the ATC condition received 26.57 ($SD = 17.82$) comparison trials throughout learning including 16.47 ($SD = 7.15$) different category combinations. This works out to one comparison trial per 9.65 trials invested, very similar to the proportion used in the NAC condition of one comparison trial for every 9.00 trials invested. Once a given category-combination was shown in a comparison trial, participants in the ATC condition confused the

same two categories again an average of 0.54 times (SD = 0.37) throughout the remainder of the learning phase.

A chi-square test of independence conducted on the ranked frequencies of category combinations demonstrated that agreement among participants was reliably different from zero, X^2 (230, N = 30) = 1026.89, $p < .001$; however, the resulting coefficient of concordance and average Spearman's correlation between participant rankings indicated that this agreement was weak, W = 0.15, r_s = 0.12.

Discussion of Preliminary ATC Work

We tested potential benefits of novel adaptively-triggered comparisons in two experiments on facial identification, in the context of a basic adaptive learning system that guides category spacing in ways previously shown to outperform non-adaptive spacing (Mettler & Kellman, 2014; Mettler et al., 2016). Experiment 1 showed that adding comparison trials contingent on individual learner confusions improved learning relative to single-item classification. Experiment 2 showed that this benefit was specific to the *adaptive* nature of the comparisons; learning was robustly enhanced (moderate to large effect sizes) by adaptively-triggered comparisons relative to similar numbers of comparison trials whose content was not adapted to individual learners. Specifically, inclusion of ATC trials led to improved learning efficiency – mastery in shorter time – than conditions without comparison trials or with non-adaptive comparisons. This held true regardless of whether efficiency was calculated with respect to the total number of trials invested or the number of minutes invested in learning. Additionally, Experiment 2 also revealed some evidence that ATCs may promote better transfer to novel instances of a category.

Notably, in the ATC conditions, the particular category pairs that triggered comparisons varied greatly across participants. Concordance tests showed very weak consistency across

participants in individual confusions between pairs of facial identity categories. Across the two experiments, 212 of 231 possible (unordered) comparisons were triggered, and only two specific comparisons occurred for 50% or more of participants (one at 56% and another at 50%). Category similarity plays a meaningful role in determining the best ways to structure learning (e.g., Carvalho & Goldstone, 2014). The weak concordance in triggered comparisons among learners indicates that similarity was highly variable across learners and underscores the value of adaptive learning technology in creating individualized opportunities for comparison.

Also noteworthy is that participants did not often recurrently trigger the same comparisons in the course of learning; any given comparison trial was repeated on average 0.70 times for a participant in Experiment 1, and 0.54 times in Experiment 2. This outcome suggests that comparison trials were highly effective in allowing participants to discover distinguishing information and avoid future confusions.

The benefit of ATC trials may be particularly impressive considering that all learning conditions also contained a highly effective adaptive learning system. Prior research suggests that on trials containing only a single exemplar, participants may make comparisons between the currently presented item and the most recently viewed item (Carvalho & Goldstone, 2017; Kang & Pashler, 2012). The way the ARTS system used here implements spacing and interleaving (Mettler et al., 2016), missed items recur over shorter intervals, and exemplars from different categories are interleaved. As a result, commonly missed targets occur in close temporal proximity to many other items. These features alone may effectively promote between-category comparisons between unmastered categories without simultaneous comparison trials of any sort. The superiority of the ATC condition, then, indicates additional value from considering an individual learner's data about missed targets and the incorrect answers given, and using this

information to structure more effective comparison opportunities.

This work tested comparisons in the domain of facial identity. Although the experimental questions were not intended to be specific to face perception, the use of faces as categories may involve special considerations. Face categories are characterized by relatively low within-category variability that may make transfer to novel instances relatively easy. Additionally, while the specific faces used in the present study were novel to all participants, adult humans may in general be considered experts in face perception and already know where to look for critical information. As a result of this learning history, facial identity categories may have reduced between-category similarity. If so, we might expect the benefits of ATCs to be even greater in other learning domains with higher between-category similarity. To explore this hypothesis, we decided to test the use of ATCs in the learning of skin lesion classifications in Experiment 5.

Experiment 5

Method

Participants

Undergraduate psychology students were recruited through the University of California, Los Angeles subject pool. We retained and analyzed the data from 67 participants. Participants were awarded partial course credit for their participation.

Materials

Stimuli consisted of dermoscopic images of the same 10 skin lesion categories used in Experiments 2-4. The number of instances available varied per category with each category containing anywhere from 19 to 96 unique images, for a total of 426 available items. Four items per category were set aside to be used as novel items in the classification assessment.

Design & Procedure

Participants were randomly assigned to either the ATC, NAC, or No Comparisons condition. These conditions were largely similar to those used in the preliminary face learning work, with the exception of a few changes in comparison triggering and frequency. Given the smaller number of categories used in the present work relative to the ATC-Faces study (10 vs. 22) and the lower learning accuracy in this domain (see Chapter 3 discussion), we modified the triggering criteria in the present work such that three relevant learning errors were required for the triggering of an ATC trial, as opposed to the two required in the prior work. Additionally, subsequent comparison trials for a specific category pairings required an additional two errors rather than one. This allowed us to better ensure that comparison trials were being triggered primarily as the result of a confusion between categories as opposed to random response behavior early in learning.

To calculate the ratio of comparison trials to single-item classification trials in the NAC condition, we used pilot data from the new ATC condition. It was estimated that participants in the ATC condition in this domain would trigger approximately one comparison trial for every 5 single-item classification trials. Using this, the NAC condition was designed to include a comparison trial on every sixth learning trial. In all conditions, categories were adaptively scheduled and interleaved through the ARTS system. In the present study, we used an enforced delay of 3 trials.

Participants continued learning trials until all categories reached mastery. Mastery criteria required four of the five last consecutive presentations of a category to be correctly classified, each given in under fifteen seconds. Once a category was mastered, it was retired from the learning set. Retired categories only re-emerged when necessary as filler items to achieve correct spacing intervals for the remaining, unmastered categories. If a participant did not reach mastery

for all 10 categories at the end of 75 minutes they were automatically progressed to the immediate posttest.

Immediately following learning, participants completed a 40-item posttest including four novel exemplars per category, randomized and presented sequentially. Posttest trials were the same as single-classification learning trials. A delayed posttest, administered one week later, was identical in content and structure to that of the immediate posttest.

Exclusion Criteria

Only data from participants who completed all parts of the experiment (pretest, learning phase, immediate posttest, and delayed posttest) were included in the following analyses. Participants who scored 30% or greater on the pretest were disqualified. Participants were excluded after data collection if they logged 15 cumulative minutes of inactivity, defined as any time not spent selecting an answer or viewing feedback after a trial. Eleven participants were excluded for this reason (ATC: $n = 2$, No Comparisons: $n = 3$; NAC: $n = 6$). Finally, one additional exclusion was made for a participant who recorded a number of learning trials greater than 3 standard deviations above the group mean (ATC: $n = 1$). The remaining total participants for analyses were as follows: ATC: $n = 23$; No Comparisons: $n = 22$; NAC: $n = 22$.

Dependent Measures & Data Analysis

We retained the same dependent measures and followed the same data analysis plan from the preliminary ATC study.

Results

Learning Investments

Learning investments were measured in terms of trials and time (minutes) invested. In total, eight participants did not reach mastery criteria for all 10 categories by the end of the 75-

minute training period. These included 1 participant in the ATC condition, 2 participants in the No Comparisons condition, and 5 participants in the NAC condition.

The number of trials invested, whether to achieve mastery of all categories or until the maximum time was reached, was lowest in the ATC condition (M = 257.04, SD = 88.22) followed by the No Comparisons condition (M = 345.05, SD = 175.33) and the NAC condition (M = 359.23, SD = 197.66). Results of Levene's test for equality of variances indicated unequal variances across conditions, $F(2, 64) = 7.40$, $p = .001$, thus we opted to use Welch's ANOVA to test for differences among means. A one-way Welch's ANOVA indicated a reliable difference between conditions in terms of trials invested, $F(2, 36.62) = 3.92$, $p = .029$, $\eta_p^2 = 0.08$. Specifically, the difference between the ATC condition and NAC condition was found to be reliable, $t(28.77) = -2.22$, $p = .034$, with a medium effect size, $d = 0.67$, as well as the difference between the ATC and No Comparisons condition, $t(30.69) = -2.11$, $p = .043$, $d = 0.63$. The No Comparisons and NAC conditions did not reliably differ, $t(41.41) = -.25$, $p = .802$, $d = 0.08$.

The number of minutes spent in the learning phase, whether to achieve mastery of all categories or until the maximum time was reached, was numerically lowest in the ATC condition (M = 36.65 min, SD = 15.85) followed by the No Comparisons condition (M = 46.01, SD = 18.79) and the NAC condition (M = 50.12, SD = 22.12). The results of Levene's test indicated unequal variance among conditions, $F(2, 64) = 3.76$, $p = .029$. A one-way Welch's ANOVA revealed that the difference among conditions boarded on reliability, $F(2, 41.54) = 3.20$, $p = .051$, $\eta_p^2 = 0.06$. Independent t-tests, with equal variance not assumed, determined the difference between the ATC and No Comparisons condition to be marginal, $t(41.12) = -1.80$, $p = .079$, $d = 0.53$, and there was no reliable between the No Comparisons and NAC conditions, $t(40.93) =$ –

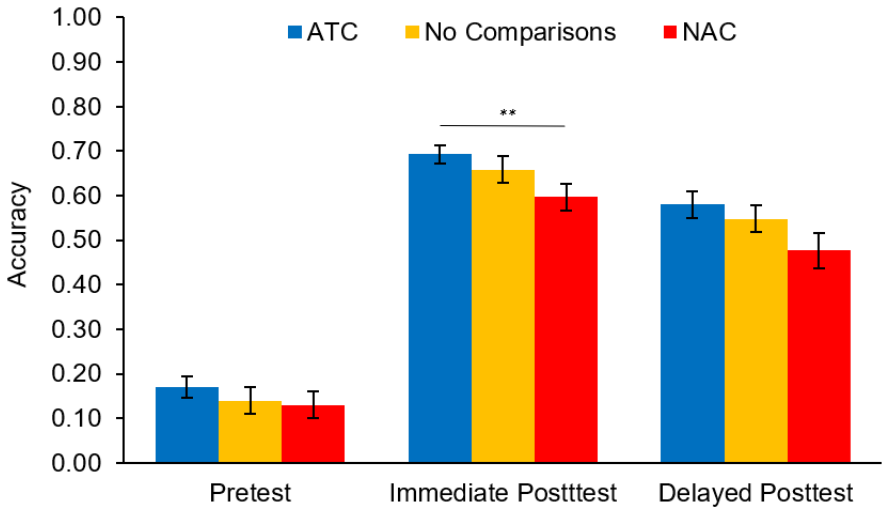
0.66, p = .510, d = 0.20. A reliable difference was found between the ATC and NAC conditions, t(37.96) = -2.34, p = .025 with a medium effect size, d = 0.70.

Accuracy Measures

Figure 21 shows assessment accuracy for each condition. Accuracy, measured as proportion correct, was not found to differ among groups at the pretest, with an overall average score of 0.15 (SD = .07), F(2, 64) = 2.37, p = .102. Posttest accuracy was compared between conditions at both the immediate and delayed posttests. At the immediate posttest, accuracy favored the ATC condition (M = .69, SD = .11) followed by the No Comparisons condition (M = .66, SD = .13) and the NAC condition (M = .59, SD = .13). Levene's test was passed for all accuracy measures (all p > .150), and subsequent analyses assumed equal variances. A one-way ANOVA on assessment scores determined the difference in immediate posttest accuracies to be reliable, F(2, 64) = 3.66, p = .031, η_p^2 = 0.10. Subsequent contrasts revealed a significant difference in favor of the ATC condition relative to the NAC condition, t(43) = 2.76, p = .009, with a large effect size, d = 0.83. The difference between the ATC and No Comparisons condition was not significant, t(43) = 1.00, p = .319, d = 0.25, nor was the difference between the No Comparisons and NAC conditions, t(42) = 1.60, p = .118, d = 0.54**.**

At the delayed posttest, no significant differences were found between conditions, F(2, 64) = 2.40, p = .099, η_p^2 = 0.07. Numerically, mean accuracy followed the same pattern as the immediate posttest, with participants in the ATC condition scoring highest (M = .58, SD = .14) followed by the No Comparisons condition (M = .55, SD = .16) and then NAC condition (M = .48, SD = .18).

Figure 21
Assessment Accuracy (Experiment 5)



Note. Accuracy, measured as proportion correct, for each of the three learning conditions at each assessment phase. Error bars indicate +/- 1 standard error of the mean. Statistical significance is indicated with asterisks: * $p < .05$, ** $p < .01$, ***$p < .001$.

Efficiency

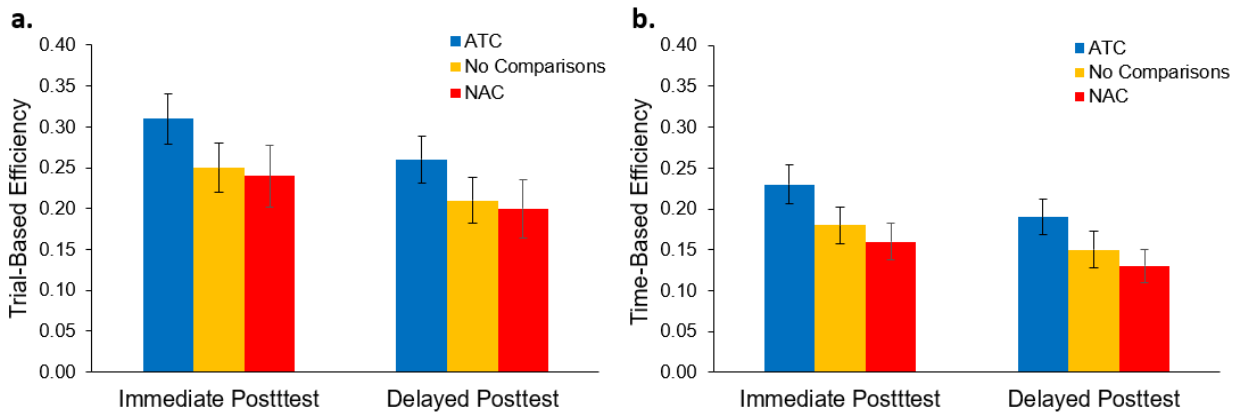
Trial-Based Efficiency

Efficiency measures were calculated for each participant at both assessment phases with mean efficiencies compared across conditions. Figure 22a shows the average trial-based efficiency by condition and posttest phase. At the immediate posttest, trial-based efficiency calculations revealed that participants in the ATC condition saw a 31% increase in immediate posttest accuracy for every 100 trials invested in learning ($M = .0031$, $SD = .0015$). In the No Comparisons condition, there was an average increase of 25% per 100 trials invested ($M = .0025$, $SD = .0014$), and a 24% increase in the NAC condition ($M = .0024$, $SD = .0018$). A one-way ANOVA determined these differences to be unreliable, $F(2, 64) = 1.19$, $p = .311$, $\eta_p^2 = 0.04$.

At the delayed posttest, participants in the ATC condition experienced a 26% increase per 100 trials invested ($M = .0026$, $SD = .0014$), participants in the No Comparisons condition experienced a 21% increase ($M = .0021$, $SD = .0013$), and participants in the NAC condition

experienced a 20% increase (*M* = .0020, *SD* = .0017). A one-way ANOVA determined these differences to be unreliable, $F(2, 64) = 0.869$, $p = .424$, $\eta_p^2 = 0.03$.

Figure 22
Efficiency Results (Experiment 5)



Note. Immediate and delayed posttest efficiency scores are shown for the Adaptively Triggered Comparisons (ATC) condition, No Comparisons condition, and the Non-Adaptive Comparisons (NAC) condition. Trial-based efficiency scores (a) indicate mean accuracy gain per 100 learning trials invested. Time-based efficiency scores (b) indicate mean accuracy gain per 10 minutes invested. Error bars represent +/- 1 standard error of the mean.

Time-Based Efficiency

Figure 22b shows the average time-based efficiency by condition and posttest phase. Time-based efficiency calculations revealed that for every 10 minutes invested in learning, the average immediate posttest score increased by 23% for participants in the ATC condition (*M* = .023, *SD* = .012). In the No Comparisons condition, there was an average increase of 18% per 10 minutes invested (*M* = .018, *SD* = .011), and a 16% increase in the NAC condition (*M* = .016, *SD* = .010). A one-way ANOVA (equal variances assumed) determined these differences to be unreliable, $F(2, 64) = 2.51$, $p = .089$, $\eta_p^2 = 0.07$.

At the delayed posttest, participants in the ATC condition experienced a 19% increase per 10 minutes invested (*M* = .019, *SD* = .011), participants in the No Comparisons condition

experienced a 15% increase (M = .015, SD = .010), and participants in the NAC condition experienced a 13% increase (M = .013, SD = .010). A one-way ANOVA determined these differences in efficiency to be unreliable, $F(2, 64) = 2.00$, $p = .144$, $\eta_p^2 = 0.06$.

Adaptive Comparisons

On average, participants in the ATC condition triggered 41.61 (SD = 21.10) comparison trials throughout learning including 19.39 (SD = 5.34) different category combinations. Once a given category-combination was shown in a comparison trial, participants in the ATC condition triggered another comparison trial between the same two categories an average of 1.04 times (SD = 0.53) throughout the remainder of the learning phase.

The average ratio of comparison trials to total trials invested was one comparison trial for every 6.49 trials–a very similar ratio to the one used in the NAC condition of 1 trial for every 6 total trials invested.

A chi-square test of independence conducted on the ranked frequencies of category combinations demonstrated that agreement among participants was reliably different from zero, X^2 (44, N = 23) = 469.73, $p < .001$. The resulting coefficient of concordance and average Spearman's correlation between participant rankings indicated that this agreement was moderately strong, $W = .464$, $r_s = 0.44$

Discussion

We tested an adaptive approach to comparisons, previously shown to enhance the learning of face categories, in the domain of skin lesion differentiation. An adaptively triggered comparison (ATC) condition was compared to a single-item No Comparisons condition, as well as a condition that leveraged the advantages of both single-item trials and comparison trials but was non-adaptive in the timing and content of the comparison trials (NAC condition). All

conditions were tested in the context of a basic adaptive learning system that guided category spacing in ways that have been previously shown to outperform non-adaptive spacing (Mettler & Kellman, 2014; Mettler et al., 2016).

Results revealed a reliable advantage of ATC learning in terms of learning investments required to achieve mastery of all categories. In particular, participants in the ATC condition reached the mastery criteria more quickly than those in the No Comparisons and NAC conditions when learning investments were quantified as the total number of learning trials invested. When learning investments were measured in terms of minutes invested, the ATC advantage was maintained over the NAC condition, but not over the No Comparisons condition. Additionally, some evidence of an ATC advantage for learning outcomes was observed at the immediate posttest, specifically with participants in the ATC condition having demonstrated an enhanced ability to classify novel exemplars relative to those who learned with non-adaptive comparisons.

The comparison trials used in the present study allowed learners to search for distinguishing information in the categories presented without needing to rely on memory of past presentations. By reducing cognitive load through the simultaneous presentation of a limited number of items (2), participants could devote more attention to extracting perceptual patterns and invariants on these trials. The superiority of the ATC condition over the NAC condition in the number of trials and minutes needed to achieve mastery of categories in training, as well as in the resulting ability to classify novel exemplars immediately following training, provides some evidence that tailoring comparisons to individual participant needs is particularly advantageous. Specifically, the adaptivity in the ATC condition may have enhanced learning in two primary ways: by selectively choosing which categories to be paired together and by presenting these trials when they were needed most.

The agreement among participants regarding which categories were most confusable was moderately strong in the present work. Two category combinations were triggered by 100% of participants in the ATC condition (Basal Cell Carcinoma-Seborrheic Keratosis and Squamous Cell Carcinoma–Wart), and four combinations were never triggered (Actinic Keratosis-Nodular Melanoma, Haemangioma-Solar Lentigo, Nodular Melanoma-Solar Lentigo, Solar Lentigo-Squamous Cell Carcinoma). The other 39 category combinations ranged from being triggered in 9% of participants to 87%. Relative to faces, the observed agreement among participants regarding the most confusable category pairings was notably higher, though still indicates considerable variation.

Despite the ATC condition yielding numerically superior performance in terms of learning investments and learning outcomes, subsequent efficiency analyses were shown not to differ reliably. There are a few possible explanations for why this was the case. First, the amount of variability in the present experiment, even within each condition, was large. Looking at the immediate time-based efficiency scores, on one extreme, one participant in the NAC condition yielded an efficiency of only 0.0049, suggesting that their performance on the immediate posttest could be expected to increase by only 5% for every 10 minutes spent in learning. Conversely, on the other extreme, the most efficient participant in this condition yielded an efficiency of 0.042, suggesting that posttest accuracy could be expected to reach 42% after only 10 minutes of learning. Similarly large ranges in efficiency can be seen in the ATC and No Comparisons condition as well. This variation is likely, in part, due to variation in participant categorization abilities; however, given that the experiment was administered online, there may also have been variations in participant behavior and attention that were not directly observed. Regardless of the source of variation, statistical significance is more difficult to achieve in the presence of high

levels of variability, and future work may aim to replicate this work in a more controlled environment to cut down on possible extraneous noise.

A second possibility concerns the frequency of the ATC trials themselves. These trials were designed such that after three relevant learning errors were committed, one exemplar from each of the confused categories was randomly chosen for presentation on the comparison trial. We increased the triggering criteria from 2 errors used in the prior work to 3 errors to address the difficulty of the domain and the decrease in set size. It is possible that by making the triggering criteria more difficult, we missed opportunities to mitigate confusions earlier on in learning that could have improved subsequent learning efficiency. Alternatively, even with this adjustment, the ratio of comparison trials to single-item trials was approximately 48% greater in the present work relative to in the face domain. It is possible that there is a balance between the number of single-item trials and the number of comparisons trials a learner receives that is most effective and this increase in comparison trials may have been disruptive. Future work may systematically test how adjustments to comparison triggering may influence the efficacy of the system.

Interestingly, we did not find any differences between learning with non-adaptive comparisons relative to learning without any comparison trials on any measures. Although the results from Chapter 3 Experiment 2 suggest that learning through (non-adaptive) paired comparisons can be very effective in learning these skin lesion categories, it is important to consider the differences that exist between learning consisting of only comparison trials and learning with only a limited number of comparison trials. As previously discussed, comparisons are most effective when the items being compared are similar (e.g., Dwyer & Vladeanu, 2009). In all paired comparisons learning, all combinations of paired comparisons were presented to participants multiple times each, thus ensuring that the most similar pairings would be included

at multiple timepoints. Meanwhile, the comparison trials in the NAC condition were limited, such that a participant performing at the mean in the NAC condition would receive less than 60 total comparison trials. It is likely that many of these trials could have contained categories that were not perceived as highly similar to the participant and/or occurred at a point in learning when a simultaneous presentation was no longer as beneficial. A worthy follow up may be to investigate how adaptively triggered comparisons, which are limited in frequency, but sensitive to perceived similarity, may compare to learning with all paired comparison trials to gain additional insight as to how one may be able to leverage advantages of both learning types.

Connections to Representational Change (Experiment 4)

In Chapter 4, Experiment 4, we measured perceived similarity between all pairwise combinations of skin lesion categories. Given the importance of differentiation in learning categories, one could reason that the categories that appear most similar would also be the most difficult to learn as they require the discovery and selection of more subtle distinguishing features. Correlation analyses did not reveal a relationship between learning performance and the similarity scores of different-category pairings; however, we theorized that if we looked at performance on an individual pairing level (e.g., similarity for the Nodular Melanoma-Wart pairing) rather than at the category level (e.g., similarity for all trials that contain Nodular Melanoma), we might see some relationship between performance and similarity emerge.

In the present study, we analyzed learning performance data (classification errors) to determine which category pairings were most confusable, and consequently, the most difficult to learn. To evaluate if there is a meaningful relationship between perceived similarity of categories and learning performance on those categories, we conducted a correlation analysis between the mean rank of confusions for category pairings in the ATC condition and the similarity data

collected in Experiment 4. In particular, we predicted that categories that were most confusable would also be the categories that yielded the highest similarity ratings. Results revealed that for each category pairing, there was a reliable relationship between the similarity ratings obtained before training (Exp. 4) and the average confusability ranking (Exp. 5), $r(43) = .311$, $p = .038$, such that for each category pairing, the higher the average confusability ranking (with a high number indicating the most confusable pairing), the more similar items from those categories were rated. A similar relationship was found between the posttest-similarity ratings and confusability rankings, $r(43) = .414$, $p = .005$, as well as between the similarity difference scores and the confusability rankings, $r(43) = .322$, $p = .031$.

The results of these correlation analyses indicate that the categories perceived to be most similar to each other, both before and after categorization training, are the same categories that were most frequently confused for each other in the present work. In addition, the way in which similarity changed as a result of training was also related to the confusability of the pairing– namely the items that failed to show evidence of between-category expansion were also triggered for comparison more frequently. These analyses provide a clear link between the perceived similarity of different-category pairings and the learning performance for those categories. This may be viewed as complementary to the relationship observed in Experiment 4 between the similarity ratings for same-category pairings and posttest performance. Taken together, a clear link can be established between the strength of category acquisition and the magnitude of resulting representational change.

Conclusion

Adaptive learning methods have been shown to be effective in enhancing the learning of perceptual classifications (e.g., Mettler & Kellman, 2014). In earlier work, we demonstrated that

using a novel adaptive comparison procedure guided by each learner's performance could enhance the learning of a large number of face categories. When applied to the difficult domain of skin lesion interpretation, we only partially replicated the results of previous work. Learning with adaptively triggered comparisons was numerically best on all measures (learning investments, posttest accuracy, and efficiency), but reliable differences were found only for learning investments and immediate posttest accuracy. While the results of this study show promise for the potential of comparison interventions in this domain, additional modification of the system may be necessary to see a significant advantage in learning efficiency.

CHAPTER 6

Summary and Concluding Remarks

Categorical learning is important and often challenging in specialized domains, such as medical image interpretation, and commonplace ones, such as face recognition. The work in this dissertation has embraced the crucial question of how to effectively prepare learners to learn the classifications of large numbers of natural, perceptual categories. Extensive research has indicated that direct comparison of items can play a meaningful role in facilitating category acquisition and generalization. Across five different experiments, my work has examined the significant relationship between the perceptual processes underlying categorization and the role of comparisons in enhancing them, ultimately offering valuable insights into how comparisons can be used to accelerate learning and improve learning outcomes.

In Chapter 3, I tested the effectiveness of paired comparisons for learning to recognize and identify the faces of 22 individuals (Experiment 1) and for learning the differential diagnosis of 10 dermatological skin lesion categories (Experiment 2). A Paired Comparisons condition, in which a learner was presented on each trial with a category label and required to choose between instances from two different categories, was compared to Single-Classification and Dual-Classification conditions, where instances of one or two categories were presented for classification on each trial. In both experiments, participants advanced through the paired comparison trials quicker than either classification-based trial format. On the assessments, the results indicated no differences in classification accuracy among any of three learning conditions in Experiment 1. However, the results of Experiment 2 showed that learning based exclusively on paired comparison trials produced greater accuracy than either classification-based condition both immediately following training, as well as after a one week delay.

The results from Experiments 1 and 2 indicate that training with paired comparison trials is an effective way to promote the learning of multi-category classifications in complex, naturalistic domains. The results of Experiment 2 in particular challenge the notion that comparison exists as an independent learning mechanism or that any simultaneous presentation of items will promote strong comparison. Paired comparison learning is effective because it positions the differentiation of categories as the goal of each trial, providing direct context for perceptual learning to occur. This, in turn, supports the ultimate goal of differentiation between all learned categories in the future. Although learning still progressed through learning trials that did not share this task, the results from the present work suggest that their improvements were weaker or slower to emerge. Therefore, paired comparison learning may be particularly advantageous in domains novel to the learner or when distinguishing information is subtle or complex.

Given the novelty of using paired comparisons to learn a large number of perceptual classifications, as well as observed effectiveness of this format, the paired comparison format was studied and tested more extensively across two experiments in Chapter 4. In Experiment 3, I broke down the paired comparison format into its separate "target" and "distractor" components to measure how learning advanced for each category on a trial-by-trial basis. To accomplish this, the frequency with which a given category occupied the target position was manipulated across three learning conditions: Always-Never, where half of the categories were always shown as target and the other half never shown as target; Often-Rarely, where half of categories appeared 75% as targets and 25% as distractors, and Equal Split learning, in which all categories appeared as targets or distractors equally often. After learning, transfer results indicated that all conditions

yielded an equivalent overall ability to generalize to novel exemplars, but categories prioritized more often as targets exhibited greater learning gains.

Why does this asymmetry emerge? We theorize that the framing of the specific categories on a paired comparison trial results in direct differences in the extraction of distinguishing features, such that a learning experience that frames a given category as the target of a trial will preferentially benefit learning to distinguish exemplars from the target category more so than from the distractor category. In other words, the task of a paired comparisons trial is perhaps more specific than simply to differentiate between the presented categories, but rather the task is to find the information that allows you to be able to differentiate the *target* category specifically. This idea builds directly off of the findings of Experiments 1 and 2, once again demonstrating that improvements in perception are directly invoked by task demands.

In Experiment 4, emphasis was shifted away from evaluating paired comparison's effects on classification accuracy, and instead aimed to uncover the underlying perceptual changes that allow for learning to occur. By measuring similarity of items from same and different category pairings before and after a comparison-based training session, representational change was observed in the difficult skin lesion domain. In particular, items from the same category came to be viewed as more similar to one another (acquired equivalence) and items from different categories came to be viewed as more dissimilar to another (acquired distinctiveness). These findings highlight the capacity of this contrast-focused learning approach to not only enhance the discrimination between different categories but also aid in the discovery of commonalities within members of the same category. Furthermore, the clear relationship between learning performance and magnitude of representational change that was observed provides additional evidence for a perceptual system that systematically updates in response to a problem or task.

Finally, while Experiments 1-4 explored how to structure comparisons and position categories most effectively on a global level, the introduction of adaptive learning methods in Chapter 5 allowed us to focus on the individual learner and consider the role of adaptive technology in advancing perceptual learning. In earlier work, we developed and tested a new type of adaptive learning element designed to leverage the benefits of learning with comparisons, termed adaptively triggered comparisons (ATCs). ATCs prompt comparison trials based on patterns of errors where a learner confuses two categories, considering not only the accuracy but also the specific incorrect responses.

Our previous research demonstrated that this adaptive comparison procedure could enhance the learning of a large set of face categories relative to learning with only single-item trials or learning with non-adaptive comparisons. In Experiment 5, the ATC procedure was tested in the domain of skin lesion interpretation. An advantage of ATC learning was observed on some measures, including learning investments needed to achieve mastery of items, but differences in overall efficiency were not found to be reliable. Adaptive learning, and in particular adaptively triggered comparisons, show promise for improving the learning of perceptual classifications by creating events in learning that identify where changes and improvements in the perception and discrimination of specific categories are most crucial. Suggestions are given as to how this system may be altered to strengthen its effects in the skin lesion domain.

Though the primary focus of this dissertation was to evaluate the role of comparisons in learning, it also provides valuable insights as to the role of perceptual learning interventions in complex, real-world tasks more generally. Experts in a given domain are often differentiated from novices by their ability to quickly identify and extract perceptual information, patterns, and

relationships. These abilities are the direct result of perceptual learning. Yet, discussions surrounding training in high-level domains, such as in STEM in medical education, are dominated by declarative and procedural learning, often overlooking perceptual learning entirely. The emphasis on declarative and procedure instruction across many domains and settings likely stems from both a lack of familiarity with perceptual learning, as well as the previous absence of effective methods to systematically induce it.

Here, we demonstrated that participants could not only learn the classification of real skin lesion images but also build effective category representations and generalize to new examples without ever being exposed to declarative information about the categories. Moreover, this training was relatively short, often occurring over the course of just one training session, and with some participants showing considerable acquisition of these categories in as little as 15 minutes. The results obtained across these experiments, particularly in Experiments 2-5, vividly illustrate the power and potential of intentional perceptual learning interventions in high-level domains. Together, the work reported here offers novel insights into how different comparison-focused manipulations can meaningfully affect learning efficacy and establishes a clear foundation for future investigation.

Appendix A

Experiment 2: Unadjusted Accuracy Analyses

A 3 (*condition*) X 2 (*posttest phase*) mixed measures ANOVA was performed on the unadjusted posttest accuracy scores. There was a reliable within-subjects effect of test phase, such that participants performed better on the immediate posttest than the delayed, $F(1, 87) = 54.41$, $p < .001$, $\eta_p^2 = 0.39$, as well as a reliable between-subjects effect of learning condition, $F(2, 87) = 6.66$, $p = .002$, $\eta_p^2 = .13$. We found no reliable evidence of an interaction between posttest phase and condition, $F(2, 87) = 0.06$, $p = .943$.

Planned pairwise contrasts between learning conditions were conducted at each posttest phase. At the immediate posttest, an independent samples t-test revealed that the accuracy advantage of the Paired Comparisons condition was significant relative to both the Dual-Classification condition, $t(58) = 2.89$, $p = .005$, $d = 0.74$, and the Single-Classification conditions, $t(58) = 2.78$, $p = .007$, $d = 0.72$. There was not a reliable difference between the Dual-Classification and Single-Classification conditions, $t(58) = 0.18$, $p = .856$, $d = 0.05$. This pattern was consistent at the delayed posttest, with a significant difference found between the Paired Comparisons condition and the Dual-Classification condition, $t(58) = 2.45$, $p = .018$, $d = 0.63$, as well as between the Paired Comparisons condition and Single-Classification condition, $t(58) = 2.42$, $p = .019$, $d = 0.63$. There was no reliable difference between the Dual-Classification and Single-Classification conditions, $t(58) = 0.18$, $p = .805$, $d = 0.07$.

References

Ahissar, M., & Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends in cognitive sciences, 8*(10), 457-464. https://doi.org/10.1016/j.tics.2004.08.011

American Cancer Society. (2023a, January 12). *Key statistics for basal and squamous cell skin cancers*. Retrieved June 19, 2023, from https://www.cancer.org/cancer/types/basal-and-squamous-cell-skin-cancer/about/key-statistics.html

American Cancer Society. (2023b, January 12). *Key statistics for melanoma skin cancer.* Retrieved June 19, 2023, from https://www.cancer.org/cancer/types/melanoma-skin-cancer/about/key-statistics.html

Andrews, J. K., Livingston, K. R., & Kurtz, K. J. (2011). Category learning in the context of co-presented items. *Cognitive Processing*, *12*(2), 161–175. https://doi.org/10.1007/s10339-010-0377-5

Ashby, F. G., & Maddox, W. T. (1998). Stimulus Categorization. In M. H. Birnbaum (Ed.), *Measurement, Judgment and Decision Making* (pp. 251–301). Academic Press. https://doi.org/10.1016/B978-012099975-0.50006-3

Ashby, F. G., & Maddox, W. T. (2005). Human Category Learning. *Annual Review of Psychology*, *56*, 149–178. https://doi.org/10.1146/annurev.psych.56.091103.070217

Ashby, S. R., Chaloupka, B., & Zeithamova, D. (2023). Category bias in similarity ratings: The influence of perceptual and strategic biases in similarity judgments of faces. *Frontiers in Cognition*, *2*. https://doi.org/10.3389/fcogn.2023.1270519

Atkinson, R. C. (1974). A*daptive instructional systems: Some attempts to optimize the learning process.* Stanford University, Institute for Mathematical Studies in the Social Sciences.

Bennett, R. G., & Westheimer, G. (1991). The effect of training on visual alignment discrimination and grating resolution. *Perception & Psychophysics*, *49*(6), 541–546. https://doi.org/10.3758/BF03212188

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. *Psychology and the real world: Essays illustrating fundamental contributions to society*, *2*(59-68).

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. *From learning processes to cognitive processes: Essays in honor of William K. Estes*, *2*, 35-67.

Carvalho, P. F., & Goldstone, R. L. (2014). Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study. *Memory & Cognition*, *42*(3), 481–495. https://doi.org/10.3758/s13421-013-0371-0

Carvalho, P. F., & Goldstone, R. L. (2015a). The benefits of interleaved and blocked study: Different tasks benefit from different schedules of study. *Psychonomic Bulletin & Review*, *22*(1), 281–288. https://doi.org/10.3758/s13423-014-0676-4.

Carvalho, P. F., & Goldstone, R. L. (2015b). What you learn is more than what you see: What can sequencing effects tell us about inductive category learning? *Frontiers in Psychology*, *6*, 130375. https://doi:10.3389/fpsyg.2015.00505

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*(1), 55–81. https://doi.org/10.1016/0010-0285(73)90004-2

Corrow, S. L., Davies-Thompson, J., Fletcher, K., Hills, C., Corrow, J. C., & Barton, J. J. S. (2019). Training face perception in developmental prosopagnosia through perceptual learning. *Neuropsychologia*, *134*, 107196. https://doi.org/10.1016/j.neuropsychologia.2019.107196

Davies-Thompson, J., Fletcher, K., Hills, C., Pancaroglu, R., Corrow, S. L., & Barton, J. J. (2017). Perceptual learning of faces: A rehabilitative study of acquired prosopagnosia. *Journal of Cognitive Neuroscience*, *29*(3), 573-591. https://doi.org/10.1162/jocn_a_01063

Dosher, B. A., Jeter, P., Liu, J., & Lu, Z. L. (2013). An integrated reweighting theory of perceptual learning. *Proceedings of the National Academy of Sciences*, *110*(33), 13678-13683.

Dosher, B. A., & Lu, Z.-L. (1999). Mechanisms of perceptual learning. *Vision Research*, *39*(19), 3197–3221. https://doi.org/10.1016/S0042-6989(99)00059-0

Dwyer, D. M., & Vladeanu, M. (2009). Perceptual learning in face processing: Comparison facilitates face recognition. *Quarterly Journal of Experimental Psychology*, *62*(10), 2055–2067. https://doi.org/10.1080/17470210802661736

Evered, A., Walker, D., Watt, A. A., & Perham, N. (2014). Untutored discrimination training on paired cell images influences visual learning in cytopathology. *Cancer Cytopathology*, *122*(3), 200–210. https://doi.org/10.1002/cncy.21370

Folstein, J. R., Gauthier, I., & Palmeri, T. J. (2012). How category learning affects object representations: Not all morphspaces stretch alike. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(4), 807–820. https://doi.org/10.1037/a0025836

Folstein, J. R., Palmeri, T. J., & Gauthier, I. (2013). Category Learning Increases Discriminability of Relevant Object Dimensions in Visual Cortex. *Cerebral Cortex*, *23*(4), 814–823. https://doi.org/10.1093/cercor/bhs067

Folstein, J. R., Palmeri, T. J., & Gauthier, I. (2014). Perceptual advantage for category-relevant perceptual dimensions: The case of shape and motion. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.01394

Folstein, J. R., Palmeri, T. J., Van Gulick, A. E., & Gauthier, I. (2015). Category Learning Stretches Neural Representations in Visual Cortex. Current Directions in Psychological Science, 24(1), 17-23. https://doi.org/10.1177/0963721414550707

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist*, *52*(1), 45–56. https://doi.org/10.1037/0003-066X.52.1.45

Gibson, E. J. (1969). *Principles of perceptual learning and development*. Appleton-Century-Crofts.

Gibson, E., Gibson, J. J., Pick, A. D., & Osser, H. (1962). A developmental study of the discrimination of letter-like forms. *Journal of Comparative and Physiological Psychology*, *55*(6), 897–906. https://doi.org/10.1037/h0043190

Gilbert, C. D., Li, W., & Piech, V. (2009). Perceptual learning and adult cortical plasticity. *The Journal of Physiology*, *587*(12), 2743-2751.

Gold, J. M., Sekuler, A. B., & Bennett, P. J. (2004). Characterizing perceptual learning with external noise. *Cognitive Science*, *28*(2), 167–207. https://doi.org/10.1207/s15516709cog2802_3

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, *123*(2), 178–200. https://doi.org/10.1037/0096-3445.123.2.178

Goldstone, R. L. (2000). Unitization during category learning. *Journal of Experimental Psychology: Human Perception and Performance, 26*(1), 86–112. https://doi.org/10.1037/0096-1523.26.1.86

Goldstone, R. L., Day, S., & Son, J. Y. (2010). Comparison. In B. Glatzeder, V. Goel, & A. Müller (Eds.), *Towards a Theory of Thinking: Building Blocks for a Conceptual Framework* (pp. 103–121). Springer. https://doi.org/10.1007/978-3-642-03129-8_7

Goldstone, R. L., Lippa, Y., & Shiffrin, R. M. (2001). Altering object representations through category learning. *Cognition*, *78*(1), 27–43. https://doi.org/10.1016/S0010-0277(00)00099-8

Goodfellow, I., Bengio, Y. & Courvile, A. (2016). *Deep learning*. Cambridge, MA: MIT Press.

Guy, G. P., & Ekwueme, D. U. (2011). Years of potential life lost and indirect costs of melanoma and non-melanoma skin cancer: A systematic review of the literature.

PharmacoEconomics, 29(10), 863–874. https://doi.org/10.2165/11589300-000000000-00000

Goren, C. C., Sarty, M., & Wu, P. Y. (1975). Visual following and pattern discrimination of face-like stimuli by newborn infants. *Pediatrics*, *56*(4), 544–549.

Guégan, S., Steichen, O., & Soria, A. (2021). Literature review of perceptual learning modules in medical education: What can we conclude regarding dermatology? *Annales de Dermatologie et de Vénéréologie*, *148*(1), 16–22. https://doi.org/10.1016/j.annder.2020.01.023

Hammer, R., Bar-Hillel, A., Hertz, T., Weinshall, D., & Hochstein, S. (2008). Comparison processes in category learning: From theory to behavior. *Brain Research*, *1225*, 102–118. https://doi.org/10.1016/j.brainres.2008.04.079

Higgins, E. J., & Ross, B. H. (2011). Comparisons in category learning: How best to compare for what. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), P*roceedings of the 33rd Annual Conference of the Cognitive Science Society*, 1388–1393.

Hock, H. S., Webb, E., & Cavedo, L. C. (1987). Perceptual learning in visual category acquisition. *Memory & Cognition*, *15*(6), 544–556. https://doi.org/10.3758/BF03198389

Homa, D. (1984). On the Nature of Categories. In G. H. Bower (Ed.), *Psychology of Learning and Motivation* (Vol. 18, pp. 49–94). Academic Press. https://doi.org/10.1016/S0079-7421(08)60359-X

Homa, D., & Chambliss, D. (1975). The relative contributions of common and distinctive information on the abstraction from ill-defined categories. *Journal of Experimental Psychology: Human Learning and Memory, 1*(4), 351–359. https://doi.org/10.1037/0278-7393.1.4.351

Homa, D., Dunbar, S., & Nohre, L. (1991). Instance frequency, categorization, and the modulating effect of experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*(3), 444–458. https://doi.org/10.1037/0278-7393.17.3.444

Homa, D., Powell, D., & Ferguson, R. (2014). Array Training in a Categorization Task. *Quarterly Journal of Experimental Psychology, 67*(1), 45–59. https://doi.org/10.1080/17470218.2013.790909

Homa, D., Proulx, M. J., & Blair, M. (2008). The Modulating Influence of Category Size on the Classification of Exception Patterns. *Quarterly Journal of Experimental Psychology, 61*(3), 425–443. https://doi.org/10.1080/17470210701238883

Hussain, Z., Sekuler, A. B., & Bennett, P. J. (2009). Perceptual learning modifies inversion effects for faces and textures. *Vision Research*, *49*(18), 2273–2284. https://doi.org/10.1016/j.visres.2009.06.014

Islami, F., Ward, E. M., Sung, H., Cronin, K. A., Tangka, F. K., Sherman, R. L., ... & Benard, V. B. (2021). Annual report to the nation on the status of cancer, part 1: national cancer statistics. *JNCI: Journal of the National Cancer Institute*, *113*(12), 1648-1669. https://doi.org/10.1093/jnci/djab131

Jacobson, & Halle, M. (1956). *Fundamentals of language.* Mouton & Co.

Jacoby, V. L., Massey, C. M., & Kellman, P. J. (2021). Enhancing perceptual learning through adaptive comparisons. *Journal of Vision, 21*(9), 2845-2845.

Jee, B. D., Uttal, D. H., Gentner, D., Manduca, C., Shipley, T. F., & Sageman, B. (2013). Finding faults: Analogical comparison supports spatial concept learning in geoscience. *Cognitive Processing*, *14*(2), 175–187. https://doi.org/10.1007/s10339-013-0551-7

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition*, *121*(3), 313–323. https://doi.org/10.1016/j.cognition.2011.08.001

Kang, S. H. K., & Pashler, H. (2012). Learning Painting Styles: Spacing is Advantageous when it Promotes Discriminative Contrast. *Applied Cognitive Psychology*, *26*(1), 97–103. https://doi.org/10.1002/acp.1801

Kao, S.-Y. Z., Ekwueme, D. U., Holman, D. M., Rim, S. H., Thomas, C. C., & Saraiya, M. (2023). Economic burden of skin cancer treatment in the USA: An analysis of the Medical Expenditure Panel Survey Data, 2012–2018. *Cancer Causes & Control*, *34*(3), 205–212. https://doi.org/10.1007/s10552-022-01644-0

Karpicke, J. D., & Roediger, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, *57*(2), 151–162. https://doi.org/10.1016/j.jml.2006.09.004

Kellman, P. J. (2002). Perceptual Learning. In *Stevens' Handbook of Experimental Psychology*. American Cancer Society. https://doi.org/10.1002/0471214426.pas0307

Kellman, P. J., & Garrigan, P. (2009). Perceptual learning and human expertise. *Physics of Life Reviews*, *6*(2), 53–84. https://doi.org/10.1016/j.plrev.2008.12.001

Kellman, P. J., & Massey, C. M. (2013). Perceptual Learning, Cognition, and Expertise. In *Psychology of Learning and Motivation* (Vol. 58, pp. 117–165). Elsevier. https://doi.org/10.1016/B978-0-12-407237-4.00004-9

Kellman, P. J., Jacoby, V., Massey, C., & Krasne, S. (2022). Perceptual Learning, Adaptive Learning, and Gamification: Educational Technologies for Pattern Recognition, Problem Solving, and Knowledge Retention in Medical Learning. In H. J. Witchel & M. W. Lee (Eds.), *Technologies in Biomedical and Life Sciences Education: Approaches and Evidence of Efficacy for Learning* (pp. 135–166). Springer International Publishing. https://doi.org/10.1007/978-3-030-95633-2_5

Kellman, P. J., Massey, C. M., Krasne, S. & Mettler, E. (2023). Connecting adaptive perceptual learning and signal detection theory in skin cancer screening. In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society*, 3251-3258.

Kellman, P. J., Massey, C. M., & Son, J. Y. (2010). Perceptual Learning Modules in Mathematics: Enhancing Students' Pattern Recognition, Structure Extraction, and Fluency. *Topics in Cognitive Science*, *2*(2), 285–305. https://doi.org/10.1111/j.1756-8765.2009.01053.x

Kendall, M. G., & Gibbons, J. D. (1990). *Rank correlation methods*. New York: Oxford University Press.

Kok, E. M., de Bruin, A. B. H., Robben, S. G. F., & van Merriënboer, J. J. G. (2013). Learning radiological appearances of diseases: Does comparison help? *Learning and Instruction*, *23*, 90–97. https://doi.org/10.1016/j.learninstruc.2012.07.004

Krasne, S., Hillman, J. D., Kellman, P. J., & Drake, T. A. (2013). Applying perceptual and adaptive learning techniques for teaching introductory histopathology. *Journal of Pathology Informatics*, *4*(1), 34. https://doi.org/10.4103/2153-3539.123991

Kurtz, K. J., & Gentner, D. (2013). Detecting anomalous features in complex stimuli: The role of structured comparison. *Journal of Experimental Psychology: Applied*, *19*(3), 219–232. https://doi.org/10.1037/a0034395

Levering, K. R., & Kurtz, K. J. (2015). Observation versus classification in supervised category learning. *Memory & Cognition*, *43*(2), 266–282. https://doi.org/10.3758/s13421-014-0458-2

Livingston, K. R., Andrews, J. K., & Harnad, S. (1998). Categorical perception effects induced by category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(3), 732–753. https://doi.org/10.1037/0278-7393.24.3.732

Mandler, J. M. (1997). Development of categorisation: Perceptual and conceptual categories. In *Infant development: Recent advances* (pp. 163–189). Psychology Press/Erlbaum (UK) Taylor & Francis.

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin*, *129*(4), 592–613. https://doi.org/10.1037/0033-2909.129.4.592

Marris, J. E., Perfors, A., Mitchell, D., Wang, W., McCusker, M. W., Lovell, T. J. H., Gibson, R. N., Gaillard, F., & Howe, P. D. L. (2023). Evaluating the effectiveness of different perceptual training methods in a difficult visual discrimination task with ultrasound images. *Cognitive Research: Principles and Implications*, *8*(1), 19. https://doi.org/10.1186/s41235-023-00467-0

Meagher, B. J., Carvalho, P. F., Goldstone, R. L., & Nosofsky, R. M. (2017). Organized simultaneous displays facilitate learning of complex natural science categories. *Psychonomic Bulletin & Review*, *24*(6), 1987–1994. https://doi.org/10.3758/s13423-017-1251-6

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207–238. https://doi.org/10.1037/0033-295X.85.3.207

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, *100*(2), 254–278. https://doi.org/10.1037/0033-295X.100.2.254

Mettler, E., & Kellman, P. J. (2014). Adaptive response-time-based category sequencing in perceptual learning. *Vision Research*, *99*, 111–123. https://doi.org/10.1016/j.visres.2013.12.009

Mettler, E., Massey, C. M., & Kellman, P. J. (2016). A comparison of adaptive and fixed schedules of practice. *Journal of Experimental Psychology: General*, *145*(7), 897–917. http://dx.doi.org/10.1037/xge0000170

Mettler, E., Massey, C., & Kellman, P. J. (2011) Improving adaptive learning technology through the use of response times. In L. Carlson, C. Holscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, 2532-2537.

Mettler, E., Massey, C. M., Burke, T., Garrigan, P., & Kellman, P. J. (2019). Enhancing adaptive learning through strategic scheduling of passive and active learning modes. In A. K. Goel, C. M. Seifert, & C. Freska (Eds.), *Proceedings of 40th Annual Conference of the Cognitive Science Society*, 768-773.

Min, R., Kose, N., & Dugelay, J.-L. (2014). KinectFaceDB: A Kinect Database for Face Recognition. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *44*(11), 1534–1548. https://doi.org/10.1109/TSMC.2014.2331215

Morton, J., & Johnson, M. H. (1991). CONSPEC and CONLERN: a two-process theory of infant face recognition. *Psychological review*, *98*(2), 164. https://doi.org/10.1037/0033-295X.98.2.164

Mundy, M. E., Honey, R. C., & Dwyer, D. M. (2007). Simultaneous presentation of similar stimuli produces perceptual learning in human picture processing. *Journal of Experimental Psychology: Animal Behavior Processes*, *33*(2), 124–138. https://doi.org/10.1037/0097-7403.33.2.124

Mundy, M. E., Honey, R. C., & Dwyer, D. M. (2009). Short Article: Superior Discrimination between Similar Stimuli after Simultaneous Exposure. *Quarterly Journal of Experimental Psychology*, *62*(1), 18–25. https://doi.org/10.1080/17470210802240614

Neisser, U. 1967. *Cognitive psychology*, New York: Appleton-Century-Crofts.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57. https://doi.org/10.1037/0096-3445.115.1.39

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(4), 700–708. https://doi.org/10.1037/0278-7393.14.4.700

Pachai, M. V., Sekuler, A. B., & Bennett, P. J. (2011). The use of horizontal information underlies face identification accuracy. *Journal of Vision*, *11*(11), 619. https://doi.org/10.1167/11.11.619

Patterson, J. D., & Kurtz, K. J. (2020). Comparison-based learning of relational categories (you'll never guess). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(5), 851–871. https://doi.org/10.1037/xlm0000758

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The Adaptive Decision Maker*. Cambridge University Press.

Petrov, A. A., Dosher, B. A., & Lu, Z.-L. (2005). The Dynamics of Perceptual Learning: An Incremental Reweighting Model. *Psychological Review*, *112*(4), 715–743. https://doi.org/10.1037/0033-295X.112.4.715

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*(3, Pt.1), 353–363. https://doi.org/10.1037/h0025953

Posner, M. I., & Keele, S. W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology, 83*(2, Pt.1), 304–308. https://doi.org/10.1037/h0028558

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*(3), 382–407. https://doi.org/10.1016/0010-0285(72)90014-X

Reppa, I., & Pothos, E. (2013). Predicting similarity change as a result of categorization. In M. Knauff, M. Pauen, N., Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, 1211-1216.

Rimoin, L., Altieri, L., Craft, N., Krasne, S., & Kellman, P. J. (2015). Training pattern recognition of skin lesion morphology, configuration, and distribution. *Journal of the American Academy of Dermatology*, *72*(3), 489–495. https://doi.org/10.1016/j.jaad.2014.11.016

Rosch, E. H. (1973). Natural categories. *Cognitive Psychology*, *4*(3), 328–350. https://doi.org/10.1016/0010-0285(73)90017-0

Rumelhart, D. E. (1989). The architecture of mind: A connectionist approach. In M. I. Posner (Ed.), *Foundations of cognitive science* (pp. 133-160). Cambridge, MA: MIT Press.

Schyns, P. G., & Murphy, G. L. (1994). The ontogeny of part representation in object concepts. In D. Medin (Ed.), *The psychology of learning and motivation* (Vol. 331, pp. 305-354). San Diego, CA: Academic Press.

Searston, R. A., & Tangen, J. M. (2017). Training perceptual experts: Feedback, labels, and contrasts. *Canadian Journal of Experimental Psychology / Revue Canadienne de Psychologie Expérimentale*, *71*(1), 32–39. https://doi.org/10.1037/cep0000124

Sha, L. Z., Toh, Y. N., Remington, R. W., & Jiang, Y. V. (2020). Perceptual learning in the identification of lung cancer in chest radiographs. *Cognitive Research: Principles and Implications*, *5*(1), 4. https://doi.org/10.1186/s41235-020-0208-x

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, *84*(2), 127–190. https://doi.org/10.1037/0033-295X.84.2.127

Sloutsky, V. M. (2010). From perceptual categories to concepts: What develops?. *Cognitive science*, *34*(7), 1244-1286. https://doi.org/10.1111/j.1551-6709.2010.01129.x

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1411–1436. https://doi.org/10.1037/0278-7393.24.6.1411

Spalding, T. L., & Ross, B. H. (1994). Comparison-based learning: effects of comparing instances during category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(6), 1251.

Sukut, S. L., D'Eon, M., Lawson, J., & Mayer, M. N. (2023). Providing comparison normal examples alongside pathologic thoracic radiographic cases can improve veterinary students' ability to identify abnormal findings or diagnose disease. *Veterinary Radiology & Ultrasound*, *64*(4), 599–604. https://doi.org/10.1111/vru.13232

Thai, K.P., Krasne, S., & Kellman, P. J. (2015). Adaptive Perceptual Learning in Electrocardiography: In D. C. Noelle, R. Dale, A. Warlaumont, S. Yoshimi, T. Matlock,

C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the Annual Conference of the Cognitive Science Society*, 2350–2355.

Thai, K., Krasne, S., & Kellman, P. (2015). Perceptual learning with adaptively-triggered comparisons. In Noelle, D. C., Dale, R., Warlaumont, A. S., Yoshimi, J., Matlock, T., Jennings, C. D., & Maglio, P. P. (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, 3000.

Roads, B. D., Xu, B., Robinson, J. K., & Tanaka, J. W. (2018). The easy-to-hard training advantage with real-world medical images. *Cognitive Research: Principles and Implications, 3*, 1-13.

Vogels, R., & Orban, G. A. (1985). The effect of practice on the oblique effect in line orientation judgments. *Vision Research*, *25*(11), 1679–1687. https://doi.org/10.1016/0042-6989(85)90140-3

Waite, S., Grigorian, A., Alexander, R. G., Macknik, S. L., Carrasco, M., Heeger, D. J., & Martinez-Conde, S. (2019). Analysis of Perceptual Expertise in Radiology – Current Knowledge and a New Perspective. *Frontiers in Human Neuroscience*, *13*. https://doi.org/10.3389/fnhum.2019.00213

Wittgenstein, L. (1953). *Philosophical investigations. Philosophische Untersuchungen* (pp. x, 232). Macmillan.